Exploring Bilingual Phoneme-Level Contrastive Learning for Personalized Dysarthric Speech Recognition

A case study

Exploring Bilingual Phoneme-Level Contrastive Learning for Personalized Dysarthric Speech Recognition

A case study

Thesis

to obtain the degree of Master of Science
at Delft University of Technology,
to be defended publicly on September 15th, 2026

by

Dorota Meszka

Faculty of Electrical Engineering, Mathematics and Computer Science,
Delft University of Technology, Delft, Netherlands,
born in Kraków, Poland.

Thesis committee:

Dr. Jorge Martinez Castaneda	Technische Universiteit Delft
Dr. Hayley Hung	Technische Universiteit Delft
MSc. Yuanyuan Zhang	Technische Universiteit Delft

CONTENTS

ABSTRACT

Dysarthric speech remains challenging for state-of-the-art automatic speech recognition (ASR) systems, despite their high accuracy on typical speech. Personalized dysarthric ASR offers a way to address the substantial differences in speech characteristics between individuals, but it is constrained by the limited amount of speaker-specific dysarthric data available for training. Speech produced by the same target speaker in another language provides an additional source of labeled data while preserving speaker-specific and dysarthric speech characteristics.

This study investigates bilingual phoneme-level contrastive learning (CL) for personalized dysarthric ASR in a case study involving a single speaker with severe dysarthria producing speech in Dutch and English. Phoneme-level modeling is used as phonemes provide a natural unit for identifying relationships between speech sounds across languages. English and Dutch speech are jointly modeled by constructing contrastive pairs from phonemes that are either shared across the two languages or manually identified as phonetically equivalent. Because these explicitly cross-lingual relationships constitute only a subset of the possible positive pairs, three positive sampling strategies and corresponding loss-weighting variants are evaluated to investigate whether giving them greater influence during training improves recognition or more strongly shapes the learned representation space. Two approaches for incorporating additional typical speech are also evaluated to investigate the effect of training-data composition.

Evaluation considers both phoneme error rate (PER) and the structure of the learned phoneme embedding space using cosine-distance, nearest-neighbor, and silhouette-based analyses. To the best of our knowledge, this is the first study to investigate bilingual phoneme-level contrastive learning for personalized dysarthric ASR.

Results show that phoneme-level CL improves recognition over bilingual fine-tuning using only Connectionist Temporal Classification (CTC), but explicitly prioritizing cross-lingual phoneme relationships does not provide a statistically significant improvement over the simpler same-symbol contrastive learning. At the representation level, bilingual and weighted objectives can bring cross-lingual phoneme representations closer together and produce better-separated phoneme clusters, but these changes do not consistently result in lower PER. Replacing dysarthric speech with typical speech significantly worsens recognition when the amount of training data is kept approximately constant, whereas adding typical speech on top of the full dysarthric training set significantly improves recognition; however, the latter condition also contains more training data, so the improvement cannot be attributed specifically to typical speech. Overall, the results show a benefit from phoneme-level contrastive learning for personalized dysarthric ASR, but no clear additional benefit from explicitly encoding canonical Dutch–English phoneme relationships. This suggests that the usefulness of bilingual contrastive learning may depend on defining cross-lingual relationships that better reflect the target speaker's actual speech patterns.

PREFACE

A little over two years ago, I arrived in the Netherlands for the first time to start my Master's degree at TU Delft. Today, at last, this journey is coming to an end. At the time, completing this degree and writing these final words felt impossibly far away. It is strange to think that two years have passed since then, and even stranger to think about how many things have changed since the moment I first arrived here.

Looking back, on that warm September day that I first arrived, I was completely unaware of what this degree would entail. Those two years haven't always been easy – in fact, rarely have they been. However, I can say with utmost certainty that these past twenty-four months have shaped me as a person in a way no other experience ever could, and I know that the lessons I learned throughout the time – both academic and personal – will stay with me forever. I hope that this thesis, with all the work, uncertainty, and persistence that went into it, can be proof of that.

Thank you to Dr. Jorge Martinez Castaneda, my supervisor, for his guidance, encouragement, and invaluable feedback throughout this project. I would also like to thank Yuanyuan Zhang, my daily supervisor, for always finding time to discuss my work with me, for her patience with my many questions, and for her support throughout the entire process.

Thank you to Prof. dr. Odette Scharenborg for sharing her expertise and helping me validate the English–Dutch phoneme equivalences that formed an important part of this work. I would also like to thank Dr. Hayley Hung for kindly agreeing to be part of my thesis committee.

Most of all, thank you to my Mom and Dad, who have shown me support every single day of this journey, and believed in me, even at times when I couldn't do so myself. And to all the other loved ones who have shown support in every way, big or small – thank you. I really mean it when I say I couldn't have done this without you.

Lastly, even though it might come off as a little too sentimental, thank you to my 22-year-old self for her relentless optimism that led her to move halfway across Europe without knowing anyone here. I hope that, as much as I sometimes curse myself for it, I never lose that part of you.

Dorota Meszka
Delft, September 2026

LIST OF SYMBOLS

Symbol	Meaning
ASR and model notation	
x	Input speech waveform.
y	Reference or target phoneme sequence.
y_i	Phoneme label of the i-th phoneme occurrence.
L	Number of phonemes in the reference sequence, $y = (y_1, \ldots, y_L)$.
$\hat{y}$	Predicted output sequence.
$P(y \mid x)$	Conditional probability of output sequence y given input waveform x.
$\mathscr{L}_{\text{CTC}}$	Connectionist Temporal Classification loss.
$\mathscr{L}_{\text{baseline}}$	Training objective of the baseline model; equal to $\mathscr{L}_{\text{CTC}}$.
t	Encoder frame index.
π_t	CTC output label at frame t.
$P(\pi_t \mid x)$	Frame-level probability of CTC label π_t given input waveform x.
$\mathbf{U}$	Sequence of representations produced by the pretrained wav2vec encoder.
$\mathbf{H}$	Sequence of hidden representations produced by the task-specific encoder.
$\mathbf{O}$	Logits/scores over the CTC output vocabulary.
$\mathbf{h}_t$	Encoder representation at frame t.
T_{enc}	Number of encoder output frames.
Phoneme-span representation notation	
i, j	Indices of phoneme occurrences or their corresponding embeddings.
s_i	First encoder frame assigned to phoneme occurrence i.
e_i	Exclusive end frame of the span assigned to phoneme occurrence i.
$\mathbf{r}_i$	Mean-pooled encoder representation for phoneme occurrence i.
$\mathbf{z}_i$	L_2-normalized phoneme-span representation derived from $\mathbf{r}_i$.

Symbol	Meaning
Contrastive-learning notation	
a	Index of the anchor phoneme occurrence.
p	Index of the positive phoneme occurrence.
n	Index of the negative phoneme occurrence.
$d(\cdot,\cdot)$	Distance function between phoneme-span embeddings; cosine distance is used in this thesis.
m	Margin used in the triplet loss.
$\mathscr{L}_{\text{triplet}}$	Phoneme-level triplet loss.
λ_{triplet}	Weight controlling the contribution of the triplet loss to the combined training objective.
$\mathscr{L}_{\text{total}}$	Combined CTC and contrastive-learning training objective.
ℓ_i	Language label of phoneme occurrence i.
$\mathscr{E}(y_a, \ell_a)$	Set of manually defined cross-lingual phoneme equivalents for an anchor with phoneme label y_a and language label ℓ_a.
$\mathscr{P}_{\text{same}}(a)$	Set of same-symbol positive candidates for anchor a.
$\mathscr{P}_{\text{equiv}}(a)$	Set of manually defined cross-lingual different-symbol equivalent positive candidates for anchor a.
$\mathscr{P}_{\text{cross-same}}(a)$	Set of cross-lingual same-symbol positive candidates for anchor a.
$\mathscr{P}_{\text{within-same}}(a)$	Set of same-language same-symbol positive candidates for anchor a.
Weighted contrastive-loss notation	
$\mathscr{T}$	Set of sampled triplets.
$\mathscr{T}_c$	Set of sampled triplets belonging to positive category c.
c	Positive-pair or triplet category.
$\mathscr{C}_A$	Set of positive categories used in the weighted equivalence-prioritized objective.
$\mathscr{C}_B$	Set of positive categories used in the weighted hierarchical objective.
$\mathscr{L}_c$	Mean triplet loss for triplets belonging to category c.
w_{equiv}	Weight assigned to manually defined cross-lingual different-symbol equivalence positives.
w_{same}	Weight assigned to same-symbol positives in equivalence-prioritized setup.
w_{cross}	Weight assigned to cross-lingual same-symbol positives in hierarchical setup.
w_{within}	Weight assigned to same-language same-symbol positives in hierarchical setup.

Symbol	Meaning
$\mathcal{L}^{EP}_{\text{weighted}}$	Weighted contrastive loss for the equivalence-prioritized sampling strategy.
$\mathcal{L}^{H}_{\text{weighted}}$	Weighted contrastive loss for the hierarchical sampling strategy.
Evaluation notation	
S	Number of phoneme substitutions.
D	Number of phoneme deletions.
I	Number of phoneme insertions.
N	Number of reference phonemes used in the PER calculation.
$d_{\cos}(\mathbf{z}_i, \mathbf{z}_j)$	Cosine distance between phoneme-span embeddings $\mathbf{z}_i$ and $\mathbf{z}_j$.
C	Set of embedding pairs belonging to a given pair category.
$\bar{d}_C$	Mean cosine distance between embedding pairs in category C.
k	Number of nearest neighbors considered in nearest-neighbor analysis.
$n_c(i)$	Number of the k nearest neighbors of embedding i belonging to category c.
$p_c(i)$	Proportion of the k nearest neighbors of embedding i belonging to category c.
M	Number of anchor embeddings included in the nearest-neighbor analysis.
$\bar{p}_c$	Mean nearest-neighbor proportion for category c across the M anchor embeddings.
$d_{\text{intra}}(i)$	Mean distance from embedding i to the other embeddings in its own phoneme cluster.
$d_{\text{inter}}(i)$	Minimum, over all other phoneme clusters, of the mean distance from embedding i to the embeddings in that cluster.
$\text{sil}(i)$	Silhouette coefficient of embedding i.
PER_{I}	PER of System I in a paired system comparison.
PER_{II}	PER of System II in a paired system comparison.
ΔPER	Difference between two compared systems, $\text{PER}_{\text{II}} - \text{PER}_{\text{I}}$.

1

THE MOTIVATION: EXPLOITING BILINGUALITY IN PERSONALIZED DYSARTHRIC ASR

1.1. MOTIVATION

Dysarthria is a motor speech disorder caused by a range of neuro-motor conditions. The speech of patients diagnosed with dysarthria is characterized by impaired articulation, reduced respiratory control and high acoustic variability, which can substantially reduce speech intelligibility. This results in communication with others being extremely difficult, which not only brings great inconvenience to the patients, but also increases their psychological burden [1]. In the study surveying post-stroke dysarthria patients [2], it is reported that more than half of the study participants reported negative changes in self-identity resulting from their speech disorder, and that those changes result in the individuals being prone to social isolation.

The underlying conditions behind dysarthria also reduce the patients' mobility and ability to live independently. Co-occurring physical disabilities and other medical conditions that dysarthria patients suffer from cause difficulties in using keyboard-, mouse- and touch-screen-based user interfaces. Voice-controlled devices can improve the quality of life of dysarthric patients tremendously. The usage of voice-powered smart-home devices can aid them in their mobility issues, voice-based technologies can be used for speech therapy, and, for isolated patients, they can serve as a companion to battle loneliness.

However, despite this potential, commercially available ASR systems perform considerably worse on dysarthric speech than on typical speech. Reported word error rates for dysarthric speakers are 26.2%–81.8% higher than for speakers without a speech disorder [3]. This performance gap is caused by several factors. First, since mainstream ASR systems are trained primarily on large amounts of typical speech, whose acoustic and articulatory characteristics differ substantially from dysarthric speech, these models generalize poorly to the characteristics of dysarthric speech. Second, collecting dysarthric

speech is slow and burdensome, since dysarthric speakers get tired more easily than typical speakers, which limits the amount of available dysarthric data. Finally, dysarthric speech varies considerably, both between individuals – because of differences in severity, underlying condition, and individual articulatory patterns – and within a single speaker's own speech across time, context, and utterances. As a result, even models trained on dysarthric speech may struggle to generalize well across speakers.

Personalized dysarthric ASR addresses this inter-speaker variability by adapting the recognition system to the speech of a specific target user. Rather than requiring one model to represent the full range of dysarthric speech patterns, a personalized model can specialize in the acoustic and articulatory characteristics of one individual. However, personalization further intensifies the data-scarcity problem: the amount of speech available from a single speaker is even smaller than the already limited amount of dysarthric speech available across speakers. Personalized dysarthric ASR therefore requires methods that make particularly effective use of the limited speech available from the target user.

One such source is speech produced by the same person in another language. Dutch and English recordings from the same target speaker share the speaker-specific characteristics associated with their motor-speech impairment, while providing different linguistic and phonetic material. This makes speech in another language particularly interesting for personalized ASR: unlike speech from another speaker, it adds training material without introducing inter-speaker variation. Previous work supports the potential value of multilingual information for dysarthric ASR. Hernandez et al. [4] found multilingual speech models to be more effective feature extractors for dysarthric ASR than monolingual models, while Steinmetz [5] showed that incorporating second-language speech from a dysarthric individual can improve recognition. Takashima et al. [6] similarly reported improvements when combining dysarthric and typical speech across languages. These findings suggest that the target speaker's Dutch and English speech should be treated as related sources of information rather than as independent datasets.

Executing this idea requires choosing an appropriate unit of comparison across languages. Whole utterances or words cannot be compared across English and Dutch, since the two languages share almost no vocabulary; however, given similarities between English and Dutch and an extensive overlap between phonetic inventories of these two languages, phoneme-based ASR can help exploit cross-lingual similarities between these two languages. Using phoneme-based multilingual training also reduces phoneme error rates compared to monolingual training - in their work, Yusuyin et al. [7] have shown that for 8 out of 10 languages tested (English, Spanish, French, Italian, Dutch, Swedish, Turkish and Tatar), multilingual models achieved lower PER results than their monolingual counterparts. Żelasko et al. [8] show that even phonemes unique to a single language benefit from being trained jointly with data from other languages, observing improvement in per-phoneme error rates for unique phonemes in French, Spanish, Vietnamese and Mandarin.

Phoneme-level training also allows the bilingual data to be used in a way that goes beyond simply increasing the amount of training data. Individual phonemes which are phonetically similar across English and Dutch, even when conventionally transcribed with different IPA symbols, may be produced by the speaker in an acoustically similar way, and could therefore reinforce one another's representation during training if the

model is explicitly encouraged to treat them as similar. Since dysarthric speech is itself characterized by variability between different phoneme representations, encouraging more robust phoneme representations in this way can help with adapting the model to the intra-speaker variability that makes dysarthric speech so difficult to recognize in the first place.

Testing this hypothesis, however, requires a training objective that can act on phoneme-level similarity directly; standard recognition objectives, such as CTC, do not do so. Contrastive learning offers exactly such a mechanism: by explicitly defining which pairs of phoneme instances should be treated as similar (positives) and which should not (negatives), and training the model to reflect this in its representation space, contrastive learning makes it possible to determine whether phonetically similar cross-lingual phoneme pairs can be pulled closer together in the learned representation space, and whether doing so is beneficial for recognition. In this setting, phonetically related Dutch and English phonemes can be treated as positive pairs. Since these cross-lingual relationships form only a part of the available positive pairs, their effect may depend on how frequently and how strongly they contribute to the contrastive objective. The study therefore considers both the selection of cross-lingual positives and the strength of their contribution during training.

Contrastive objectives have previously shown promising results for dysarthric ASR specifically: Wu et al. [9] use contrastive learning with data-augmentation-based positive pairs to improve dysarthric recognition, Wang et al. [10] integrate supervised contrastive learning in their dysarthric ASR model, and DyPCL [11] introduces a dynamic, difficulty-ordered contrastive training scheme for dysarthric speech. None of this prior work, however, uses contrastive learning to exploit cross-lingual phonetic relationships in dysarthric speaker's own bilingual data.

This research gap motivates the investigation of bilingual phoneme-level contrastive learning for personalized dysarthric ASR. In particular, this thesis examines how explicitly exploiting cross-lingual phonetic relationships between the target speaker's Dutch and English speech through phoneme-level contrastive learning affects personalized dysarthric ASR performance and the structure of the learned phoneme representation space. To investigate this, several aspects of the contrastive learning framework are considered, including how cross-lingual positive pairs are selected, how strongly these relationships are emphasized in the training objective, and how the incorporation of additional typical speech affects the proposed approach. These aspects form the basis of the research questions presented in the following section.

1.2. RESEARCH QUESTIONS

The main research question is:

> **RQ:** How does bilingual phoneme-level contrastive learning with cross-lingual positive pairs affect personalized dysarthric ASR and the learned representation space?

To answer this question, the following sub-research questions are addressed:

SRQ1: How do different triplet sampling strategies affect recognition performance and representation structure?

This sub-question investigates whether different ways of selecting positive samples for the triplet loss affect the model's recognition performance and learned representation space. In particular, it compares basic same-symbol sampling, equivalence-prioritized bilingual sampling, and hierarchical bilingual sampling. The comparison considers both phoneme error rate and embedding-space analyses.

SRQ2: How does weighting the bilingual contrastive loss affect recognition performance and representation structure compared to the unweighted objective?

This sub-question examines whether assigning higher weights to triplets with bilingual positives changes the effect of the contrastive objective. Both phoneme error rate and embedding space analyses are conducted.

SRQ3: How does incorporating typical speech data affect recognition performance and representation structure?

This sub-question explores the role of typical speech data in the proposed framework. It compares two settings: one where typical speech replaces part of the dysarthric training data, and one where typical speech is added on top of the available dysarthric data.

1.3. THESIS OUTLINE

The thesis is organized as follows: Chapter 2 introduces background knowledge for better understanding of this thesis. Chapter 3 explains the methodological framework of this thesis. Chapter 4 describes the experimental setup, including the dataset, data pre-processing, training configurations, and evaluation metrics. Chapter 5 reports the experimental results. Chapter 6 provides the findings of these results and answers to the research questions. Chapter 7 presents the conclusions and discusses directions for future work.

2

THE FOUNDATIONS: ASR, DYSARTHRIA, AND CONTRASTIVE OBJECTIVES

2.1. AUTOMATIC SPEECH RECOGNITION

Automatic speech recognition (ASR) is the task of automatically converting a spoken utterance into a symbolic linguistic representation. In conventional ASR, the input is a speech signal, and the desired output is a sequence of words or characters corresponding to what was said [12].

A standard ASR pipeline can be viewed as estimating the most likely output sequence given an input speech signal. Let $\mathbf{x}$ denote the input audio and $\mathbf{y}$ denote the target linguistic sequence. The goal of ASR is to find the sequence $\hat{\mathbf{y}}$ that is most likely given the acoustic input:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} P(\mathbf{y} \mid \mathbf{x}).$$

2.1.1. END-TO-END ASR

Traditional ASR systems were typically composed of three subcomponents: an acoustic model, a pronunciation model or a lexicon, and a language model. The acoustic and language models were each trained separately with their own data: the acoustic model used speech utterances with transcriptions for training, and the language model was trained on text. The pronunciation model was typically constructed from a fixed pronunciation dictionary rather than trained. These components were then combined during decoding.

The introduction of end-to-end ASR models combined these components into a single, jointly optimized neural framework [13]. This simplified pipeline, compared with training and combining several separate components, has contributed to end-to-end systems becoming widespread and achieving state-of-the-art results on different ASR benchmarks [14].

Several approaches are commonly used for end-to-end ASR, the most widely used of which are Connectionist Temporal Classification (CTC)-based models, Attention-based Encoder-Decoder (AED) models, and Recurrent Neural Network Transducers (RNN-T). These approaches differ in how they model the relationship between the input speech and the output sequence. Conventional AED models use a decoder that applies an attention mechanism to assign different weights to the encoder representations when generating each output label, resulting in soft input-output alignments that do not necessarily correspond to precise phoneme boundaries. RNN-T models introduce a separate autoregressive prediction network in addition to the encoder and combine the two through a joint network, so that each output prediction is conditioned on the previously generated labels. CTC uses frame-level output predictions without requiring an autoregressive decoder.

This frame-level output structure is particularly useful for the method proposed in this thesis. The CTC output probabilities can be combined with the reference phoneme sequence in a forced alignment procedure to obtain explicit phoneme spans over the encoder representations. As described in Section 3.2, these spans are then used to extract the phoneme-level embeddings on which the contrastive objective operates. CTC therefore provides a comparatively simple recognition objective while also giving access to the frame-level structure required by the proposed phoneme-level contrastive learning framework.

2.1.2. CONNECTIONIST TEMPORAL CLASSIFICATION

Connectionist Temporal Classification (CTC) is a training objective for sequence prediction tasks in which the alignment between the input frames and the output labels is unknown [15]. CTC adds a special blank symbol, representing frames that are not assigned to any output label, to the output vocabulary and produces a probability distribution over this extended label set at each input frame.

A single output sequence can correspond to multiple frame-level prediction paths. During training, CTC therefore does not require a predetermined frame-level alignment. Instead, it defines the probability of the target sequence by summing over all valid paths that collapse to that sequence. A path is collapsed by merging repeated labels and removing blank symbols, and the required marginalization over possible paths is computed using dynamic programming.

At inference time, the frame-level output probabilities can be decoded greedily or using beam search. Since CTC does not use an autoregressive decoder, the frame-level predictions are conditionally independent given the input representation.

2.1.3. SELF-SUPERVISED SPEECH REPRESENTATION LEARNING

Modern ASR systems typically require large amounts of labeled speech data, but producing accurate transcriptions is expensive and time-consuming. This is especially problematic in low-resource settings, where labeled data may be scarce, while unlabeled speech recordings are easier to obtain. Self-supervised speech representation learning addresses this problem by learning representations from speech without requiring manually provided labels during pretraining. That way, a model can first learn general acoustic and phonetic structure from large amounts of unlabeled speech, and then reuse this

knowledge when it is fine-tuned on a much smaller labeled dataset. This reduces the amount of task-specific labeled data required to learn useful speech representations. The approach is loosely inspired by human language acquisition, where knowledge about speech develops largely through exposure to spoken language, without relying on explicit labelled supervision [16].

Self-supervised learning is related to unsupervised learning because both approaches make use of unlabeled data. However, self-supervised learning creates its own training targets from the input data. For example, a model may mask part of an audio representation and learn to identify or reconstruct the missing information from the surrounding context. This allows the model to learn structure in the speech signal before it is fine-tuned on a labeled downstream task.

Self-supervised speech representation learning includes several types of training objectives, including generative and contrastive approaches [17]. Generative approaches learn representations by reconstructing or predicting information from the input, while contrastive approaches learn by distinguishing related representations from unrelated alternatives. Despite the differences in the pretraining objective, the common goal is to learn speech representations from unlabeled data that can subsequently be transferred to downstream tasks with limited labeled data.

A prominent model family based on contrastive self-supervised speech learning is wav2vec. The original wav2vec model introduced contrastive self-supervised pretraining for speech by learning representations from raw audio [18]. Wav2vec 2.0 extended this framework by masking parts of the latent speech sequence and using a Transformer context network to learn contextualized representations [19]. During pretraining, the model learns to distinguish the correct masked latent representation from distractors. This contrastive objective differs from the phoneme-level contrastive objective introduced in Chapter 3, which explicitly structures representations according to phoneme identity and cross-lingual phonetic relationships. After pretraining, the wav2vec 2.0 encoder can be fine-tuned for ASR using labeled data.

One notable example of wav2vec extensions is XLS-R. XLS-R extends the wav2vec 2.0 framework to multilingual speech representation learning. Instead of pretraining on speech from a single language, XLS-R is pretrained on unlabeled speech from multiple languages using the same general self-supervised objective as wav2vec 2.0 [20]. This allows the model to share latent speech representations between languages and use language features from other languages without pre-training the model again in the new language [21]. This approach is particularly useful in low-resource languages with limited amount of labeled speech data [20, 22, 23].

XLS-R's multilingual pretraining is particularly relevant to the personalized dysarthric ASR setting considered in this thesis. Training a large acoustic encoder from scratch is difficult due to a small amount of data available, so using a pretrained multilingual self-supervised model allows the system to reuse representations learned from large amounts of speech and adapt them using comparatively little speaker-specific data. It is also well matched to the bilingual setup of this work, since the proposed contrastive objective operates on Dutch and English speech within a shared encoder and explicitly exploits phonetic relationships between the two languages. Because XLS-R has already been pretrained on speech from multiple languages, it serves as a good starting point for

learning shared representations across the target speaker's Dutch and English speech.

2.1.4. ASR FAIRNESS

State-of-the-art ASR systems achieve good performance on standard benchmarks, with some of them achieving almost human-like parity [24–26]. However, this performance is largely limited to typical speech – that is, speech from adult, typically highly-educated, first-language speakers of a standardized language variety, without a speech disability [27]. When tested on more diverse speaker populations, substantial performance disparities emerge.

Performance gaps in modern ASR systems have been documented across several speaker-related factors, including race [28–31], gender [28, 32, 33], age [34, 35], accent [33, 35–38] and the absence or presence of speech disorders [39, 40]. These disparities have been observed across different languages. In Dutch ASR, studies have found performance gaps for speakers with non-native accents, language proficiency, age, and speech motor impairment [41–43]. Similar disparities have also been reported for English ASR across race, gender, age, accent, and speech disorders [28, 33, 38, 39].

These gaps lead to ASR-based technologies being less accessible and less reliable, thus excluding those underrepresented speaker groups from using those systems effectively.

Similar issues are observed in ASR systems for low-resource languages. Since large-scale ASR training data mostly centers around English and other high-resource languages, systems often perform worse for languages with less available labeled speech data [44–47]. This creates an additional form of access inequality, where the benefits of modern ASR are not distributed evenly across languages and speaker communities.

2.2. MULTILINGUAL AND CROSS-LINGUAL ASR

Multilingual ASR refers to speech recognition systems that are trained on or designed to recognize speech from more than one language [48]. Instead of developing a completely separate model for each language, multilingual systems use a shared model, or shared components of a model, to learn from speech across several languages. Cross-lingual learning refers more specifically to the sharing or transfer of information between languages. A multilingual ASR system can therefore enable cross-lingual transfer when representations learned from one language are also useful for recognizing speech in another language. Multilingual ASR has become increasingly common as a way of extending speech recognition to a wider range of languages [49].

One important application of multilingual ASR is recognition in low-resource languages. When only a limited amount of labeled speech is available for a particular language, training a model using that language alone can be difficult. Multilingual training allows the model to learn from speech in several languages within a shared representation space. Since some acoustic and phonetic properties are shared across languages, this can allow the information learned from higher-resource languages to contribute to the representation of lower-resource languages, rather than treating each language independently [48]. Babu et al. [20] showed that multilingual models can learn representations that are not tied exclusively to a single language and can support cross-lingual sharing. Toshniwal et al. [50] similarly found that multilingual pretraining can improve recognition

compared with independently trained monolingual models. However, the composition of multilingual training data remains important, since an imbalance between high- and low-resource languages can negatively affect performance for some languages [20].

Phoneme-based ASR is particularly suitable for cross-lingual learning because different languages often share identical or phonetically similar speech sounds. This makes it possible for a multilingual model to learn a shared representation space in which related phonemes from different languages are represented similarly, rather than treating each language independently. Yusuyin et al. [7] showed that pooling data from multiple languages can reduce phoneme error rates compared with monolingual training. Żelasko et al. [8] further showed that multilingual training can benefit the recognition of phones in the target language, including phones that are not shared directly with the other training languages. These findings suggest that multilingual training can provide useful phonetic knowledge, even when the phoneme inventories of the languages do not fully overlap.

This distinction between multilingual training and cross-lingual sharing is particularly relevant to the approach investigated in this thesis. The ASR system is multilingual, since it is trained on both Dutch and English speech, while the proposed contrastive objective is explicitly cross-lingual because it uses phonetic relationships between Dutch and English phonemes when defining positive pairs. The aim is therefore not only to train a single model on two languages, but to investigate whether explicitly encouraging information sharing between related phonemes can improve the learned representations and recognition performance.

2.3. DYSARTHRIA AND DYSARTHRIC ASR

2.3.1. DYSARTHRIC SPEECH CHARACTERISTICS

Dysarthria is a motor speech disorder which can be caused by a range of neuro-motor conditions including cerebral palsy, amyotrophic lateral sclerosis (ALS), Parkinson disease, stroke or traumatic brain injuries [51]. The neuro-motor problems that dysarthric patients suffer from manifest themselves in weakness or paralysis of muscles that are used in articulation, which reduces the intelligibility of their speech [52].

Dysarthric speech has several pathological characteristics, such as discontinuous pronunciation, uncontrolled volume, slow speech, explosive pronunciation, improper pauses, excessive nasal sounds, and air-flow noise during pronunciation, which differ from typical speech [1]. The presence and degree of these characteristics vary across individuals. Dysarthria severity describes the overall degree of speech impairment and is commonly characterized using categories such as mild, moderate, severe, and profound, with speech intelligibility often used as one indicator of severity [53].

2.3.2. CHALLENGES OF DYSARTHRIC ASR

Dysarthric speech remains very challenging for commercially available ASR systems. Dysarthric speech tested on such ASR systems provides 26.2% - 81.8% higher word error rate (WER) compared to typical speech [3].

This performance disparity can be attributed to several aspects. First, the aforementioned acoustic differences between typical speech and dysarthric speech make it difficult for the systems mostly trained on typical speech data to adapt to dysarthric speech. Sec-

ond, collecting dysarthric speech data also poses challenges. Dysarthric speakers are easily exhausted, are less able to express emotions, and are prone to drooling and dysphagia [1]. This makes it extremely difficult to collect large quantities of disordered speech, which are essential to train any modern ASR system. Lastly, dysarthric speech has high variability. This variability occurs in two dimensions – between different speakers (known as inter-speaker variability), making it difficult for the system to perform well across all speakers, and for the speech produced by the same dysarthric speaker in different times, context or utterances (known as intra-speaker variability).

2.3.3. RECENT DIRECTIONS IN DYSARTHRIC ASR

Several approaches have been explored to address two major challenges of dysarthric ASR: the limited availability of dysarthric training data and the high acoustic variability of dysarthric speech.

One common direction is data augmentation, which addresses data scarcity by increasing the acoustic variability of the existing training data without requiring additional recordings from dysarthric speakers. By exposing the model to modified versions of the available speech, augmentation can reduce overfitting to the limited training set and encourage the recognizer to become more robust to variation in speech production. Relatively simple transformations, including speed, pitch, tempo, volume, and vocal tract length perturbations, have been shown to improve dysarthric ASR performance [54–56]. More advanced techniques such as SpecAugment [57] and domain-specific variants such as Dysarthric SpecAugment [58] have also been successfully applied.

Another way of addressing data scarcity is to generate additional training speech. Text-to-speech and voice conversion systems have been used to produce synthetic dysarthric speech that can supplement recordings from real speakers [59–61]. Rather than modifying existing recordings, these approaches attempt to create new examples that reproduce characteristics of dysarthric speech, thereby increasing the amount of training material available to the recognizer.

A second direction focuses on modifying the speech representation itself so that information relevant to recognition becomes easier for the model to distinguish. Dysarthric speech can exhibit substantial acoustic variability and atypical articulatory patterns, so the standard acoustic representation may not always make the underlying phonetic structure sufficiently clear for the recognizer. Feature-enhancement approaches therefore aim to transform or enrich the input representation before recognition. For example, articulatory features have been combined with acoustic features to provide additional information about speech production in dysarthric and other pathological speech [62]. Other work has used deep autoencoders [63] and TDNN-based denoising autoencoders [64] to enhance dysarthric speech features. Prananta et al. [65] similarly applied MaskCycleGAN-based feature modification.

While these approaches address important challenges of dysarthric ASR, particularly data scarcity and acoustic variability, they do not directly address the substantial differences in speech characteristics between individual dysarthric speakers. This motivates speaker-specific approaches that adapt the recognition system to the characteristics of an individual user.

2.3.4. PERSONALIZED DYSARTHRIC ASR

One of the main challenges of dysarthric ASR is the variability of dysarthric speech across speakers. The acoustic and articulatory characteristics of dysarthria depend on factors such as the underlying condition, severity, disease progression, and individual motor impairments. Consequently, two speakers with similar clinical characteristics can still produce speech that differs considerably, making it difficult for a single speaker-independent ASR system to perform equally well across individuals [66]. This inter-speaker variability is particularly important for speakers with severe dysarthria, whose speech may differ substantially from both typical speech and the dysarthric speech represented in the training data.

Personalized dysarthric ASR addresses this problem by adapting the recognition system to the speech characteristics of a specific target speaker. Rather than requiring one model to represent the full range of variation across dysarthric speakers, a personalized model can specialize in the acoustic and articulatory patterns of one individual. This reduces the amount of variability that the model needs to account for, making the recognition problem more manageable for that particular user.

Previous work provides strong evidence for the value of speaker-specific adaptation. Shor et al. [67] showed that personalized fine-tuning can substantially improve recognition of impaired speech, with considerable gains obtained from only a few minutes of speaker-specific training data. Green et al. [68] evaluated personalized ASR models for hundreds of speakers with disordered speech and found that personalized systems substantially outperformed speaker-independent models. The largest improvements were observed for speakers with more severe speech impairments, for whom general-purpose recognition is particularly difficult. These findings show that speaker-specific adaptation can be effective both when relatively little target-speaker data is available and across a large and diverse population of speakers.

Single-speaker studies allow the effects of speaker-specific adaptation to be examined in greater detail. Zhang et al. [69] evaluated state-of-the-art ASR systems on continuous Dutch speech from a speaker with severe dysarthria and found that general-purpose systems performed poorly, while fine-tuning on speech from the target speaker substantially improved recognition. More recently, Ogasawara et al. [70] introduced SS-JDSC, a single-speaker Japanese dysarthric speech corpus containing approximately 15 hours of speech. Fine-tuning an ASR model on this corpus substantially reduced recognition error, further showing the value of collecting sufficiently rich speaker-specific data for personalized dysarthric ASR.

Personalization nevertheless introduces an additional data challenge. While a speaker-independent dysarthric ASR system can utilize recordings from many speakers, a personalized system is restricted to the amount of speech that can be collected from its target user. Such data is scarce and hard to collect, so it is important to use the available speaker-specific data as effectively as possible.

2.3.5. MULTILINGUAL DYSARTHRIC ASR

Multilingual data provide another potential source of additional speech information for dysarthric ASR, particularly when dysarthric training data in the target language is limited. Speech produced by the same dysarthric speaker in another language still reflects many of

the speaker-specific characteristics associated with their motor-speech impairment, while also introducing additional phonetic and acoustic variation. This additional variation may therefore provide useful information for recognizing dysarthric speech. Takashima et al. [6], for example, supplemented speech from a Japanese dysarthric speaker with English dysarthric speech and typical Japanese speech. Their approach resulted in improvements of 14.7% and 6.5% for two dysarthric speakers, showing that speech outside the immediate target-speaker and target-language data can provide useful additional information for recognition.

In the study of Hernandez et al. [4], the authors found that multilingual speech models are more effective feature extractors than monolingual models for models trained to recognize monolingual dysarthric speech. The authors attribute this to multilingual pretraining exposing the model to a wider range of phonetic variation than monolingual pretraining, since the same or similar phonemes are pronounced differently across languages. They argue that this broader exposure to phonetic variation, rather than a narrower, language-specific one, makes the resulting representations better suited to the acoustic variability that characterizes dysarthric speech. Following this hypothesis, jointly using Dutch and English speech from the same dysarthric speaker may expose the model to a broader range of phonetic realizations while preserving the speaker-specific characteristics of the target dysarthric speech. If phonetic information can be shared across the two languages, this additional cross-lingual variation may help the model learn more robust phoneme representations for personalized dysarthric ASR.

In his work, Steinmetz [5] discovers that incorporating second-language speech from a dysarthric individual can improve the performance of a dysarthric ASR system, in both speaker-dependent and speaker-independent settings. The author attributes these improvements to the characteristics of second-language speech which resemble properties of dysarthric speech, such as increased variability and reduced fluency. Second-language speech may therefore provide additional acoustic and phonetic variation while, in the speaker-dependent setting, still preserving characteristics of the target speaker. This makes it a potentially useful source of additional data when only limited speech is available in the target language.

2.3.6. USING TYPICAL SPEECH FOR DYSARTHRIC ASR

Since collecting large amounts of dysarthric speech is difficult, previous work has also explored supplementing limited dysarthric training data with more readily available speech from typical speakers. Although typical speech does not exhibit the same acoustic characteristics as dysarthric speech, it can still provide additional examples of the phonetic and linguistic patterns that the ASR system needs to learn.

Yılmaz et al. [71] proposed an approach that incorporated typical Dutch speech data for Flemish pathological ASR. Their findings showed that this data merging technique in their proposed context could improve the recognition of pathological speech. In their study, Leivaditi et al. [72] found that a model fine-tuned on both dysarthric and typical speech outperformed the model fine-tuned only on dysarthric speech across all tested severity groups. However, the mixed-data system was trained on a larger amount of data, so the improvement could not be attributed specifically to the typical speech itself. Takashima et al. [6] additionally report improvements when combining Japanese

dysarthric speech with English dysarthric speech and typical Japanese speech.

These findings suggest that typical speech may be useful as supplementary training data for dysarthric ASR. At the same time, it is unclear whether the observed improvements come from the use of typical speech specifically or simply from having more training data. They also raise the question of how much typical speech can substitute for speaker-matched dysarthric speech when the total amount of training data is kept constant.

2.4. CONTRASTIVE LEARNING

2.4.1. APPROACHES TO SHARED REPRESENTATION LEARNING

Shared representation learning aims to learn features that capture information useful across different domains or conditions, while reducing differences that are irrelevant to the target task. In speech recognition, this can be useful when speech varies across factors such as language, accent, speaker, or speech condition. One approach is domain-adversarial learning, which encourages the model to learn features that are useful for the main task but less dependent on the domain of the input [73]. In speech recognition, the domain can correspond to factors such as accent, language, or speech condition [74–76].

In a domain-adversarial neural network, the encoder is optimized for the main recognition task while also being connected to a domain classifier [73]. The domain classifier attempts to determine which domain an input belongs to, while the encoder is trained to make this classification difficult. The resulting representations are encouraged to preserve information relevant to the main task while reducing domain-specific differences.

The domain is defined by the type of variation that the model is intended to become less sensitive to. In accented ASR, for example, standard and accented speech can be treated as different domains, allowing the model to learn representations that are less dependent on accent [74]. Domain-adversarial learning has also been applied to dysarthric ASR. Woszczyk et al. [76] treated typical and dysarthric speech as separate domains and trained the feature extractor to make the two less distinguishable while retaining information required for speech recognition.

The same principle can be applied across languages. In multilingual ASR, the language of an utterance can be used as the domain label. The language classifier attempts to predict the input language from the encoder representations, while the encoder is trained to reduce the amount of language-specific information they contain. Adams et al. [75] used such a language-adversarial objective in multilingual ASR to encourage language-independent encoder representations. In this way, domain-adversarial learning can encourage the model to learn representations that are more similar across languages.

However, domain-adversarial learning treats each domain as a single category and does not explicitly model finer-grained relationships within or across domains. Making Dutch and English representations less distinguishable, for example, does not explicitly determine which particular phonemes in the two languages should have similar representations. Contrastive learning provides more direct control over these individual relationships. By defining particular examples as similar or dissimilar, the objective can explicitly encourage selected representations to become closer or farther apart. This property is particularly relevant to the approach investigated in this thesis, where specific

phonetic relationships between Dutch and English phonemes are used to structure the shared representation space.

2.4.2. CONTRASTIVE LEARNING RATIONALE

Contrastive learning is a representation learning approach based on comparisons between similar and dissimilar samples. A positive pair consists of samples that should have similar representations, while a negative pair consists of samples that should remain distinguishable [77]. Training encourages representations of positive samples to become closer in the embedding space and representations of negative samples to become farther apart.

Contrastive learning can be used in both unsupervised and supervised settings. In unsupervised contrastive learning, explicit class labels are not used to define similarity between samples. Positive pairs are instead typically constructed from different views or augmentations of the same input, while other samples are treated as negatives [78]. In supervised contrastive learning, label information is available and can be used to define these relationships directly: samples belonging to the same class are treated as positives, while samples from different classes are treated as negatives [79]. The definition of these positive and negative relationships determines what structure the model is encouraged to learn.

Contrastive learning was first adapted in the computer vision domain [78–82] and has since been applied across a wide range of domains [83–86], including speech.

2.4.3. TRIPLET LOSS

In order to be able to utilize the contrastive learning objective in model training, a loss that enforces the relationships between the anchor, the positive sample(s) and the negative sample(s) needs to be defined. Several loss formulations can be used for contrastive learning, including batch-based objectives that compare an anchor against multiple positive and negative examples. In this thesis, triplet loss [87] is used, which operates on one anchor, one positive sample, and one negative sample. This formulation provides a straightforward way to deliberately select different types of positive relationships, including sparse manually defined cross-lingual equivalents, and later weight the resulting triplet categories separately.

Given an anchor sample $\mathbf{z}_a$, a positive sample $\mathbf{z}_p$, and a negative sample $\mathbf{z}_n$, the triplet loss encourages the distance between the anchor and positive to be smaller than the distance between the anchor and negative by at least a margin m:

$$\mathscr{L}_{\text{triplet}} = \max\left(d(\mathbf{z}_a, \mathbf{z}_p) - d(\mathbf{z}_a, \mathbf{z}_n) + m, 0\right),$$

where $d(\cdot,\cdot)$ is a distance function. If the negative sample is already sufficiently farther away from the anchor than the positive sample, the loss becomes zero. Otherwise, the model is penalized and the embeddings are updated so that the positive pair becomes closer and the negative pair becomes more separated.

The margin controls how much farther the negative sample should be from the anchor compared with the positive sample. A larger margin enforces stronger separation between positive and negative pairs, while a smaller margin allows the model to satisfy the constraint more easily.

Treating all triplets equally assumes that every positive and negative relationship is equally informative for shaping the representation space. In practice, however, some triplets may provide a stronger or more meaningful training signal than others. Weighting the triplet loss allows the model to reflect these differences by assigning greater influence to selected relationships. This can be useful when some positives are more strongly related than others, or when certain triplets are more difficult and therefore more informative during training. Song et al. [88], for example, propose a Softmax weighting strategy that assigns greater weight to harder samples based on the relative distances between the anchor, positive, and negative representations. Xiong et al. [89] similarly use a weighted triplet loss in which selected positive samples receive greater importance. Weighted triplet loss therefore provides a way of encoding not only which samples should be treated as related, but also how strongly those relationships should influence the learned representation space.

2.4.4. POSITIVE SAMPLING

Positive sampling determines which examples are treated as similar in a contrastive objective. In unsupervised contrastive learning, positive pairs are commonly created by applying different data augmentations to the same input, while other examples in the mini-batch are treated as negatives [90].

In supervised contrastive learning, positive examples can instead be defined using label information. Samples belonging to the same target class are typically treated as positives, while samples from different classes are treated as negatives. Restricting positive pairs to exact class identity can, however, be limiting when only a small number of examples of each class occur within a mini-batch [91]. It also prevents the objective from representing relationships between samples that belong to different classes but are nevertheless similar according to other information.

Several approaches therefore extend positive sampling beyond exact class identity. Wang et al. [91], for example, distinguish between strong positives belonging to the same class and weak positives that are similar according to additional criteria. They further weigh these relationships differently, allowing more closely related samples to contribute more strongly to the objective. Cabannes et al. [92] and Sobal et al. [93] similarly investigate contrastive objectives in which similarity is informed by relationships beyond exact label identity. These approaches show that positive sampling can incorporate known relationships between different classes, rather than treating all cross-class pairs as unrelated. Weighting provides an additional mechanism for expressing the strength of these relationships when some positive pairs are expected to be more informative or more closely related than others.

2.4.5. CONTRASTIVE LEARNING FOR SPEECH

Speech provides several forms of similarity that can be exploited by contrastive learning. Different speech segments can represent the same phoneme, multiple transformed versions can originate from the same utterance, and speech sounds can be related according to their acoustic or phonetic properties. Contrastive objectives can therefore be used to encourage representations to preserve these relationships.

Early applications of contrastive learning to speech include Contrastive Predictive

Coding (CPC) [94] and wav2vec [18], which use contrastive objectives to learn speech representations from large amounts of unlabeled data. Contrastive learning has also been combined directly with supervised ASR objectives. Talnikar et al. [95], for example, jointly trained CTC and contrastive objectives and showed that the additional representation-learning objective can improve downstream speech recognition.

Contrastive learning has also been applied in settings where recognition data is limited, including low-resource language ASR [96, 97]. When only a small number of labeled examples are available, encouraging the same or related speech units to have similar representations can help the model learn a more stable structure from the available data rather than treating every observation independently.

This property is also relevant to dysarthric ASR, where limited training data is combined with substantial acoustic variability. Contrastive learning can encourage acoustically different realizations of the same underlying speech unit to remain close in the representation space, while preserving separation between different units. Wu et al. [9] proposed a contrastive learning framework in which different data augmentation techniques were used to construct positive pairs, allowing them to investigate which transformations produced useful invariances for dysarthric speech recognition. Wang et al. [10] also incorporated supervised contrastive learning into a dysarthric ASR system.

More recently, Lee et al. introduced DyPCL [11], a phoneme-level contrastive learning approach in which the difficulty of the contrastive relationships changes during training. Negative samples are ordered according to phonetic similarity, so that easier distinctions are introduced before increasingly similar phonemes. By incorporating phonetic similarity into the construction of the contrastive task, the method encourages the model to learn increasingly fine-grained distinctions between related speech sounds.

The studies above show that the definition of similarity is an important design choice in contrastive learning for speech. In particular, the choice of which samples are treated as positives or negatives determines which relationships the model is encouraged to preserve in the learned representation space.

2.5. PHONEME-BASED ASR

2.5.1. PHONEMES AS RECOGNITION UNITS

Phonemes are the smallest structural units of speech that distinguish meaning between words [98]. Phoneme sounds are broken into four broad classes: vowels, diphthongs, semivowels and consonants. While the specific phoneme inventory differs from language to language, many phonemes are shared across languages, and the International Phonetic Alphabet (IPA) [99] provides a standardized notation system for representing speech sounds, making it possible to describe and compare the phonetic inventories of different languages.

Due to a small inventory size compared to word vocabularies, phonemes have historically been used as an intermediate recognition unit in ASR. By recognizing phoneme sequences and mapping them to text through a pronunciation lexicon, these models could generalize to unseen words built from known phonemes.

2.5.2. WHY PHONEME-BASED ASR?

Phoneme inventories often overlap between languages, so using phoneme as a basic unit for ASR systems allows for cross-lingual mapping in multilingual ASR systems [98]. According to [7], especially in limited data settings, phoneme-based supervised pre-training achieves the most competitive results. In contrast, the methods using grapheme units face challenges in learning shared cross-lingual representations due to a lack of shared graphemes among different languages [100]. As presented in Section 2.2, in phoneme-based multilingual ASR, introducing multilingual data greatly benefits the models.

In the multilingual contrastive learning setup that aims to benefit from cross-lingual similarities, using phonemes as a representative unit is the only way to do so. Except for loan-words, the vocabularies of two languages are different, so utterance- or word-level samples are impossible to compare to each other. For phonemes however, determining similarity is possible, and it allows to build mappings that reflect similarities and dissimilarities between phoneme sets of two languages.

Phoneme-level modeling can also alleviate sparsity in the output space compared with word-level modeling. Since phonemes are reused across many words, each phoneme can be observed in multiple lexical contexts, providing more training examples per unit than would be available for individual word-level targets. Previous work has shown that phoneme-based representations can improve recognition in low-resource speech recognition settings [101], making them particularly relevant to personalized dysarthric ASR, where only limited amounts of target-speaker speech are available.

3

THE METHOD: BILINGUAL PHONEME-LEVEL CONTRASTIVE LEARNING

3.1. BASELINE MODEL ARCHITECTURE

The proposed contrastive learning methods described later in this chapter build on a baseline phoneme recognition model.

The baseline model consists of a pretrained wav2vec-based acoustic encoder [19, 20] followed by a task-specific encoder and a linear CTC classification layer [15]. Given an input waveform, the pretrained encoder extracts a sequence of frame-level speech representations, which are then passed through the task-specific encoder, producing frame-level hidden representations. These resulting hidden representations are then projected to the phoneme vocabulary by the linear CTC output layer. A log-softmax operation converts these scores into frame-level log-probabilities over the CTC output labels. The overall forward pass can be summarized as follows:

$$x \rightarrow \text{wav2vec}(x) = \mathbf{U} \rightarrow \text{Enc}(\mathbf{U}) = \mathbf{H} \rightarrow \text{CTCLinear}(\mathbf{H}) = \mathbf{O} \rightarrow \log \text{softmax}(\mathbf{O}) \rightarrow \log P(\pi_t \mid x),$$

where x is the input waveform, $\mathbf{U}$ represents the wav2vec encoder representations, $\mathbf{H}$ denotes the hidden representations, $\mathbf{O}$ represents logits/scores over the CTC output vocabulary, π_t denotes the CTC output label at frame t and $P(\pi_t \mid x)$ is its frame-level probability.

The model is trained using the Connectionist Temporal Classification (CTC) loss. For each utterance, CTC defines the probability of the target phoneme sequence by summing over all valid prediction paths that collapse to that sequence after removing repeated labels and blank symbols. CTC therefore allows the model to learn from phoneme sequences without requiring a fixed frame-level alignment between phonemes and acoustic frames.

The baseline objective is therefore:

$$\mathscr{L}_{\text{baseline}} = \mathscr{L}_{\text{CTC}}.$$

3.2. PHONEME-LEVEL EMBEDDING EXTRACTION

The contrastive learning objective operates on phoneme-level representations rather than on whole utterance representations. Therefore, the frame-level encoder output of the baseline model must be converted into a sequence of embeddings corresponding to individual phoneme occurrences. This requires an alignment between the reference phoneme sequence and the frame-level representations produced by the model.

To obtain these alignments, the best checkpoint of the baseline model is used. For each utterance, frame-level encoder representations and CTC log-probabilities are computed. The reference phoneme sequence is then aligned to these frame-level predictions using a CTC forced alignment procedure [102] as follows:

Given an input waveform x and its reference phoneme sequence

$$y = (y_1, \ldots, y_L),$$

the target sequence is expanded with blank symbols. Dynamic programming is then used to find the highest-scoring frame-level path corresponding to the reference sequence. From this path, each non-blank phoneme occurrence is assigned a span of encoder frames.

Each extracted span is represented by a start frame and an end frame:

$$(s_i, e_i),$$

where s_i is the first frame assigned to phoneme y_i, and e_i is the exclusive end frame of that span. Given the encoder output sequence:

$$\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_{T_{\text{enc}}}),$$

where T_{enc} denotes the number of encoder output frames, the embedding for phoneme occurrence i is obtained by mean pooling over the encoder frames inside the corresponding span:

$$\mathbf{r}_i = \frac{1}{e_i - s_i} \sum_{t=s_i}^{e_i-1} \mathbf{h}_t.$$

The resulting vector $\mathbf{r}_i$ is used as the phoneme-level representation in the contrastive learning objective.

The alignments are computed for two span extraction variants. In the first variant, the start and end frames of each phoneme span are taken directly from the non-blank regions of the CTC forced alignment path. In this case, frames assigned to the CTC blank symbol are not included in the phoneme embedding. In the second variant, the start frame of each phoneme is identified first, and the end frame is defined as the start frame of the next phoneme. This produces contiguous phoneme spans by assigning the region between two phoneme starts to the preceding phoneme. The first variant gives narrower phoneme spans, while the second avoids gaps between consecutive phoneme embeddings.

The two variants differ in how frames assigned to the CTC blank symbol are incorporated into phoneme representations. Non-blank spans exclude these frames, potentially producing more phoneme-specific embeddings but using fewer encoder frames. Contiguous spans include the intervening blank frames, which may preserve useful transitional or coarticulatory information but can also introduce information from neighboring phonemes. Both variants are evaluated to test whether span definition affects the effectiveness of phoneme-level contrastive learning.

Both span definitions are extracted once before contrastive learning and kept fixed during training. This allows the two span extraction strategies to be compared with each other.

A dynamic alignment variant was also explored, in which alignments were recomputed during training using the current model predictions. Dynamic CTC alignment has also been used in prior phoneme-level contrastive learning work, such as DyPCL [11]. However, since the dynamic variant did not lead to noticeable differences in the results, the main experiments used precomputed fixed alignments. The results for experiments using dynamic alignment can be found in Section A.3 of the Appendix.

The mechanisms of both variants are illustrated in Figure 3.1.

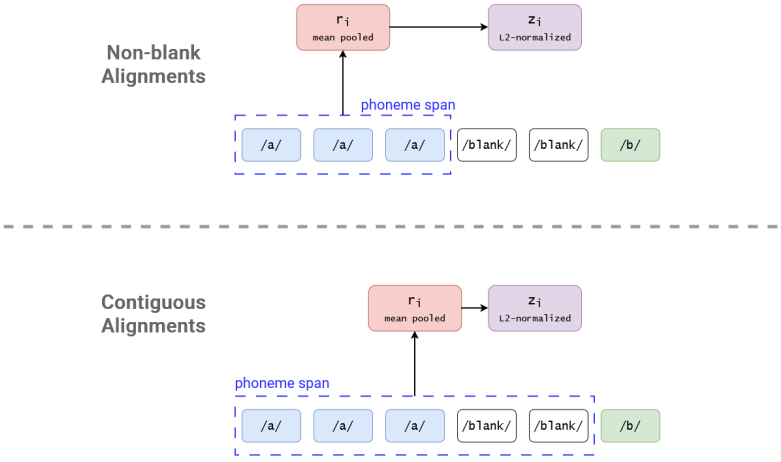

Figure 3.1: Phoneme-span embedding extraction under the two alignment variants. Non-blank spans (top) use only the frames assigned to the phoneme by the CTC forced-alignment path, excluding blank-labeled frames, producing narrower spans that may leave gaps between consecutive phonemes. Contiguous spans (bottom) instead extend each phoneme's span to the start frame of the next phoneme, assigning all intervening frames, including those labeled blank, to the preceding phoneme. In both variants, frames within the span are mean-pooled into r_i and L_2-normalized into z_i, the embedding used in the contrastive objective.

3.3. PHONEME-LEVEL CONTRASTIVE LEARNING

The baseline CTC model is extended with a phoneme-level contrastive learning objective.

The contrastive loss is applied to the phoneme-span representations extracted from the encoder output, as described in Section 3.2. Since cosine distance is used by the triplet loss, the representations are normalized to ensure that similarity is determined by their direction rather than their magnitude:

$$\mathbf{z}_i = \frac{\mathbf{r}_i}{\|\mathbf{r}_i\|_2}$$

The contrastive objective is then computed directly on the normalized encoder representations. A projection head was also considered as an additional transformation before the contrastive objective, following common practice in contrastive representation learning [78], but it was not used in the main experiments, since the early experiments performed better without it.

The contrastive objective is formulated as a triplet loss [87]. For each anchor phoneme embedding $\mathbf{z}_a$, a positive phoneme embedding $\mathbf{z}_p$ and a negative phoneme embedding $\mathbf{z}_n$ are selected. The positive is intended to represent a phoneme occurrence that should be close to the anchor, while the negative represents a phoneme occurrence that should be separated from it. The positive sampling strategies are described in Section 3.5. The negative is selected from phoneme occurrences with a different label from the anchor and positive, excluding phonemes defined as cross-lingual equivalents of the anchor.

The triplet loss is then computed using the cosine distance:

$$d(\mathbf{z}_i, \mathbf{z}_j) = 1 - \cos(\mathbf{z}_i, \mathbf{z}_j).$$

For a triplet (a, p, n), the loss is:

$$\mathcal{L}_{\text{triplet}} = \max\big(d(\mathbf{z}_a, \mathbf{z}_p) - d(\mathbf{z}_a, \mathbf{z}_n) + m, 0\big),$$

where m is the triplet margin. This objective penalizes triplets for which the negative is not at least m farther from the anchor than the positive, thereby encouraging this relative distance constraint during training.

The full training objective combines the CTC loss with the phoneme-level triplet loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CTC}} + \lambda_{\text{triplet}} \mathcal{L}_{\text{triplet}},$$

where λ_{triplet} is a hyperparameter that controls the contribution of the contrastive objective.

The contrastive loss is applied only during training. During validation and testing, the model is evaluated using the CTC phoneme predictions.

During the experiments, a variant was explored in which contrastive learning was applied only after a warm-up period of a few epochs. However, because this setup showed worse results than applying contrastive learning immediately, it was abandoned.

3.4. CROSS-LINGUAL PHONEME EQUIVALENCE MAPPING

To incorporate bilingual information into the contrastive learning objective, it was necessary to define which English and Dutch phonemes could be treated as equivalent. Without such a mapping, the contrastive objective would be limited to identical phoneme symbols

and would introduce the risk of treating phonetically different sounds as positives only because they share the same IPA notation.

The equivalence mapping was based on a comparative phonetic analysis of the two languages, using *The Phonetics of English and Dutch* [103] as the main reference. This source provides a detailed comparison of the phonetic systems of the two languages and was used to identify similarities and differences between the relevant phoneme inventories. The resulting mapping was later validated by consulting an expert in phonetics, who is also a native Dutch speaker.

The literature was used to establish the phonetic characteristics and cross-linguistic similarities of the relevant sounds; the decision to treat particular pairs as equivalent for contrastive learning is a methodological choice made in this thesis rather than a claim that the phonemes are linguistically identical.

The analysis had two goals. First, phonemes represented by the same IPA symbol in English and Dutch were examined to determine whether they could be treated as sufficiently similar for the purpose of contrastive learning. If so, occurrences of these phonemes could be used as positive pairs in the triplet loss objective. Second, the analysis aimed to identify cross-lingual phoneme pairs which are represented by different IPA symbols but are phonetically similar enough to be treated as equivalent in the bilingual contrastive learning setup.

After the analysis, the following pairs with different IPA symbols were determined to be equivalent:

English phoneme	Dutch phoneme
ɑː	aː
oʊ	eʊ
oʊ	oː
aʊ	ʌʊ
aʊ	ʉ
eɪ	eː
eɪ	ɛɪ
eɪ	ɛː
eɪ	ɪː
ɐ	ə

Table 3.1: English-Dutch different IPA symbol phoneme pairs treated as equivalent for bilingual contrastive learning.

After the analysis, it was also determined that not all phonemes with shared IPA symbols should be treated as equivalent. In particular, English and Dutch /w/ and /v/ were kept distinct, since their realizations differ sufficiently across the two languages. These phonemes were assigned language-specific labels in the implementation, preventing them from being treated as same-symbol positives during contrastive learning.

3.5. BILINGUAL POSITIVE SAMPLING STRATEGIES

The previous section defined which English-Dutch phoneme pairs may be treated as equivalent for the purposes of contrastive learning. This section describes how that equivalence information is used during triplet construction. In particular, the following sampling strategies differ in how the positive example for each anchor is selected: using only identical phoneme symbols, or using a prioritized combination of manually defined cross-lingual equivalents and same-symbol positives.

Let $\mathbf{z}_i$ denote the normalized embedding of a phoneme occurrence, y_i its phoneme label, and ℓ_i its language label. For an anchor $(\mathbf{z}_a, y_a, \ell_a)$, the positive example is selected from the other phoneme embeddings available in the same mini-batch.

Previous contrastive-learning approaches have shown that positive relationships can be defined using similarity information beyond exact class identity [91–93]. Building on this idea, the strategies considered here use cross-lingual phonetic relationships to define additional positive candidates.

3.5.1. BASIC SAME-SYMBOL SAMPLING

The basic contrastive learning strategy uses only same-symbol positives. For a given anchor phoneme occurrence, the set of positive candidates consists of all other phoneme occurrences in the batch with the same phoneme label:

$$\mathscr{P}_{\text{same}}(a) = \{\, i \neq a \mid y_i = y_a \,\}.$$

A positive example is sampled from this set when at least one same-symbol candidate is available.

This strategy does not use the manually defined English-Dutch equivalence mapping. Therefore, phonemes represented by different IPA symbols are never treated as positives, even if they are phonetically similar across the two languages. However, when an English and a Dutch phoneme share the same label in the vocabulary, they may be selected as positives for one another. This makes the strategy a simple baseline for phoneme-level contrastive learning, where similarity is defined only through identical phoneme labels.

Language-specific labels are used for phonemes that share an IPA symbol but should not be treated as equivalent across English and Dutch. In particular, English and Dutch /w/ and /v/ are represented using separate labels, preventing them from being selected as same-symbol positives.

3.5.2. EQUIVALENCE-PRIORITIZED BILINGUAL SAMPLING

The first bilingual sampling strategy extends the same-symbol sampling by incorporating the manually defined cross-lingual equivalence mapping. For each anchor, the sampler first checks whether a cross-lingual different-symbol equivalent is available in the same mini-batch. Let $\mathscr{E}(y_a, \ell_a)$ denote the set of manually defined equivalent phoneme-language pairs for anchor phoneme y_a in language ℓ_a. The equivalence-based positive candidate set is:

$$\mathscr{P}_{\text{equiv}}(a) = \{\, i \neq a \mid (y_i, \ell_i) \in \mathscr{E}(y_a, \ell_a) \,\}.$$

If $\mathscr{P}_{\text{equiv}}(a)$ is non-empty, the positive example is sampled from this set. This gives priority to cross-lingual phoneme pairs that are represented by different IPA symbols but

were manually identified as equivalent. If no such equivalent is available in the batch, the sampler falls back to same-symbol positives:

$$\mathscr{P}_{\text{same}}(a) = \{ i \neq a \mid y_i = y_a \}.$$

In this version, all same-symbol positives are treated as a single category. That is, the strategy does not distinguish between same-symbol positives from the other language and same-symbol positives from the same language. The priority order is therefore:

cross-lingual different-symbol equivalent
↓
same-symbol positive

This strategy is designed to test whether manually defined cross-lingual equivalence pairs provide an additional useful signal beyond basic same-symbol contrastive learning, while still allowing anchors to contribute to the contrastive loss when no explicit equivalent is present in the mini-batch.

Equivalence-prioritized sampling strategy is illustrated in the top panel of Figure 3.2

3.5.3. HIERARCHICAL BILINGUAL SAMPLING

The second bilingual sampling strategy further distinguishes between different types of same-symbol positives. As in the previous strategy, the sampler first checks whether a manually defined cross-lingual different-symbol equivalent is available in the mini-batch:

$$\mathscr{P}_{\text{equiv}}(a) = \{ i \neq a \mid (y_i, \ell_i) \in \mathscr{E}(y_a, \ell_a) \}.$$

If this set is non-empty, the positive is sampled from it.

If no manually defined equivalent is available, the sampler next checks for cross-lingual same-symbol positives. These are phoneme occurrences that have the same phoneme label as the anchor but come from the other language:

$$\mathscr{P}_{\text{cross-same}}(a) = \{ i \neq a \mid y_i = y_a \wedge \ell_i \neq \ell_a \}.$$

If this set is non-empty, the positive is sampled from it.

Finally, if neither a manually defined equivalent nor a cross-lingual same-symbol positive is available, the sampler falls back to same-language same-symbol positives:

$$\mathscr{P}_{\text{within-same}}(a) = \{ i \neq a \mid y_i = y_a \wedge \ell_i = \ell_a \}.$$

The motivation for this hierarchy is to prioritize positive pairs that make the strongest use of bilingual information, while still allowing the contrastive objective to be applied when such pairs are not available. Manually defined cross-lingual different-symbol equivalents are given the highest priority because they directly test the main bilingual hypothesis: that phonetically similar English and Dutch phonemes can be brought closer together even when they are represented by different IPA symbols. However, these pairs are relatively sparse and may not be present for every anchor in every mini-batch. Cross-lingual same-symbol positives are therefore used as the next-best option, since they still

encourage the bilingual objective by using both English and Dutch realizations of the same phoneme label. Finally, same-language same-symbol positives are used to preserve the basic phoneme-level contrastive signal when no cross-lingual positive is available.

The resulting priority order is:

cross-lingual different-symbol equivalent
↓
cross-lingual same-symbol positive
↓
same-language same-symbol positive

This strategy gives the strongest priority to explicitly defined bilingual phoneme relationships, while still making use of same-symbol positives when no manually defined equivalent is available. Compared with the previous strategy, it separates cross-lingual and same-language same-symbol positives, allowing the sampler to prioritize bilingual positive pairs whenever possible, fully exploiting the bilingual nature of the training data and the potential representation robustness it provides.

Hierarchical sampling strategy is illustrated in the bottom panel of Figure 3.2

3.6. BILINGUAL BATCH CONSTRUCTION

The triplet sampling strategies proposed above are heavily dependent on the availability of suitable cross-lingual positive phoneme samples in each mini-batch. Explicit cross-lingual equivalence positives can only be sampled when the corresponding English and Dutch phoneme samples appear in the same batch. Forming batches randomly introduces a risk that no usable cross-lingual equivalence pairs will be present, causing the contrastive objective to fall back to same-symbol or same-language positives and making the proposed approach ineffective. To reduce this risk, the mini-batches are computed offline in a way that increases the availability of cross-lingual positive pairs.

Each batch was constructed to contain six utterances: three English utterances and three Dutch utterances. The batch construction procedure is guided by the predefined list of phoneme equivalences, as described in Section 3.4. For each equivalence pair, the training set was searched for English utterances containing the English-side phoneme and Dutch utterances containing the Dutch-side phoneme. Equivalence pairs were processed starting from the least frequent pairs, measured by the smaller number of available English or Dutch utterances. This prioritizes rare equivalence pairs, which would otherwise be less likely to appear in randomly formed batches.

For each equivalence pair, batches were created by selecting three English utterances containing the English phoneme and three Dutch utterances containing the corresponding Dutch phoneme. Once no further complete batch could be formed for a given pair, the procedure moved to the next equivalence pair. This strategy increases the likelihood that the contrastive sampler can select explicit cross-lingual equivalence positives within a batch, while maintaining a balanced English-Dutch batch composition. Remaining utterances that could not be assigned through the equivalence-guided procedure were grouped into additional balanced batches of the same fixed size by repeatedly taking three unused English utterances and three unused Dutch utterances from the remaining

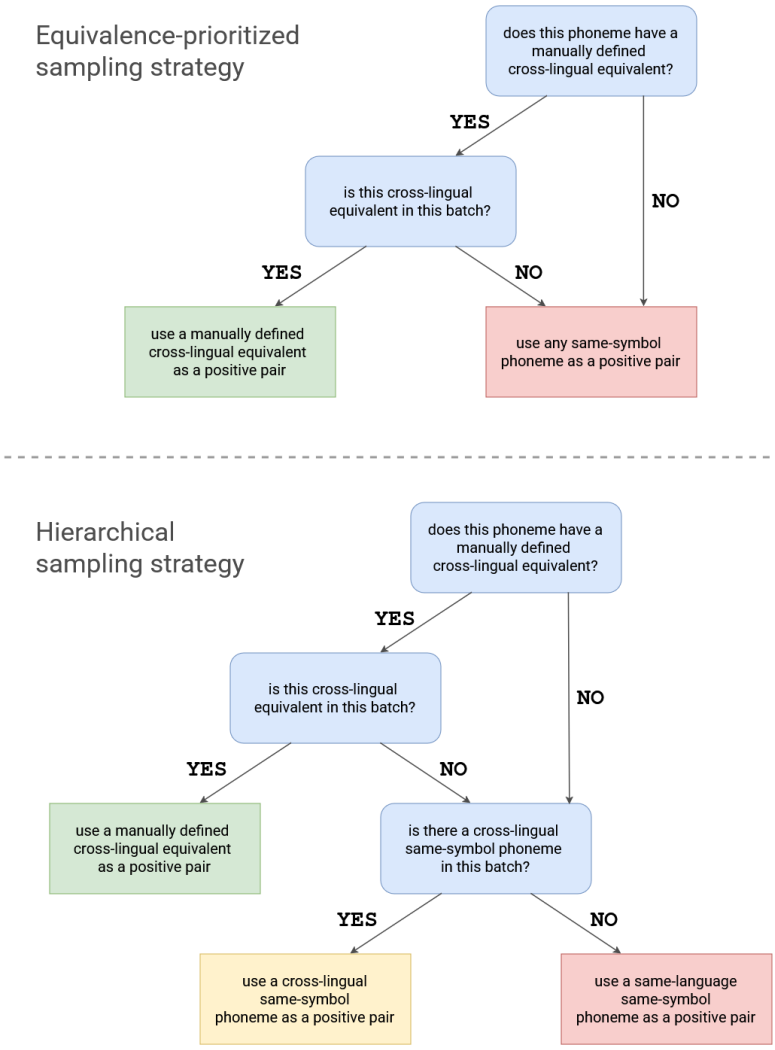

Figure 3.2: Equivalence-prioritized sampling (top) falls back to a single undivided same-symbol category when no manually defined cross-lingual equivalent is found. Hierarchical sampling (bottom) splits this further, preferring cross-lingual same-symbol positives over same-language ones.

sets, without applying any further phoneme-based selection criterion; no explicit random sampling procedure was applied.

3.7. CONTRASTIVE LOSS VARIANTS

The positive sampling strategies described in the previous section determine which positive example is selected for each anchor. However, after triplets have been constructed, it is also necessary to define how the losses from different triplet types are combined. This is particularly relevant for the bilingual sampling strategies, where some triplets are based on manually defined cross-lingual different-symbol equivalences, while others are based on same-symbol positives.

Two contrastive loss variants are considered. The first treats all sampled triplets equally. The second assigns different weights to triplets depending on the priority level of the positive example used to construct them.

3.7.1. STANDARD TRIPLET LOSS

In the standard triplet loss variant, all valid triplets are treated equally, regardless of how the positive example was selected. Given a set of sampled triplets $\mathcal{T}$, the contrastive loss is computed as the average triplet loss over all triplets:

$$\mathcal{L}_{\text{triplet}} = \frac{1}{|\mathcal{T}|} \sum_{(a,p,n)\in\mathcal{T}} \max\left(d(\mathbf{z}_a, \mathbf{z}_p) - d(\mathbf{z}_a, \mathbf{z}_n) + m, 0\right),$$

where a, p, and n denote the indices of the anchor, positive, and negative phoneme occurrences, respectively, $d(\cdot,\cdot)$ is the cosine distance, and m is the triplet margin.

This loss treats triplets based on manually defined cross-lingual equivalences and triplets based on same-symbol positives in the same way. Therefore, any difference between the sampling strategies comes only from which positives are selected, not from how strongly their corresponding losses are weighted.

3.7.2. WEIGHTED BILINGUAL TRIPLET LOSS

The weighted bilingual triplet loss extends the standard triplet loss by assigning different weights to triplets depending on the type of positive example used. Previous work has shown that weighting can be used to increase the influence of triplets that are considered more informative or to give greater importance to selected positive relationships [88, 89]. In the present setting, this idea is applied to the different cross-lingual positive categories.

The motivation is that these positive types do not have the same methodological role. Cross-lingual different-symbol equivalence positives are most directly related to the bilingual hypothesis of this thesis, but they are also relatively sparse. If all triplets are averaged together without distinction, their contribution to the overall loss may be small compared with same-symbol triplets, which occur more frequently.

The weighted loss therefore allows the training objective to place greater emphasis on triplets that more directly reflect the intended cross-lingual relationships of the bilingual contrastive learning setup. Rather than weighting individual triplets directly, the loss is first computed separately for each positive category. These category-level losses are then combined using manually defined weights.

For the equivalence-prioritized bilingual sampling strategy, there are two positive categories:

$$\mathscr{C}_A = \{\text{equiv}, \text{same}\},$$

where equiv denotes triplets whose positive example is a manually defined cross-lingual different-symbol equivalent, and same denotes triplets whose positive example is selected from the same-symbol set.

Let $\mathscr{T}_{\text{equiv}}$ and $\mathscr{T}_{\text{same}}$ denote the sets of triplets belonging to these two categories. The category-specific losses are:

$$\mathscr{L}_c = \frac{1}{|\mathscr{T}_c|} \sum_{(a,p,n)\in\mathscr{T}_c} \max\left(d(\mathbf{z}_a, \mathbf{z}_p) - d(\mathbf{z}_a, \mathbf{z}_n) + m, 0\right),$$

for each available category c. When both categories are present, the weighted equivalence-prioritized loss is:

$$\mathscr{L}^{\text{EP}}_{\text{weighted}} = \frac{w_{\text{equiv}}\mathscr{L}_{\text{equiv}} + w_{\text{same}}\mathscr{L}_{\text{same}}}{w_{\text{equiv}} + w_{\text{same}}},$$

where w_{equiv} and w_{same} control the relative contribution of manually defined equivalence positives and same-symbol positives. In this thesis, the equivalence-positive category is assigned a higher weight in order to increase the influence of the cross-lingual different-symbol triplets.

For the hierarchical bilingual sampling strategy, the weighted loss follows the three priority levels used during positive sampling. The positive categories are:

$$\mathscr{C}_B = \{\text{equiv}, \text{cross-same}, \text{within-same}\},$$

where equiv denotes manually defined cross-lingual different-symbol positives, cross-same denotes cross-lingual same-symbol positives, and within-same denotes same-language same-symbol positives. When all three categories are present, the weighted hierarchical loss is:

$$\mathscr{L}^{\text{H}}_{\text{weighted}} = \frac{w_{\text{equiv}}\mathscr{L}_{\text{equiv}} + w_{\text{cross}}\mathscr{L}_{\text{cross-same}} + w_{\text{within}}\mathscr{L}_{\text{within-same}}}{w_{\text{equiv}} + w_{\text{cross}} + w_{\text{within}}}.$$

This version allows the loss weighting to reflect the hierarchy used in positive sampling. Manually defined cross-lingual different-symbol positives receive the highest emphasis, cross-lingual same-symbol positives receive intermediate emphasis, and same-language same-symbol positives receive the lowest emphasis. This tests whether explicitly emphasizing more bilingual triplet types in the loss improves the learned phoneme representation space compared with treating all triplets equally.

If a given category is not present in a mini-batch, it is excluded from the weighted average for that batch. This prevents missing triplet categories from contributing zero-valued losses and ensures that the normalization only uses categories that are actually available.

3.8. INCORPORATING TYPICAL DATA

As discussed in Section 2.3.6, previous work has shown that typical speech can provide useful supplementary training data for dysarthric or pathological ASR [6, 71, 72]. Based on this, an exploratory set of experiments was conducted to examine whether incorporating typical speech also benefits the personalized dysarthric ASR system studied in this thesis.

However, adding typical data introduces a methodological distinction between two possible effects. Improvements may result from the presence of typical speech itself, but they may also result simply from increasing the total amount of training data. To separate these effects, two typical-data variants were considered.

The first variant uses a matched-size typical-dysarthric training condition. In this setup, the total length of training examples is kept approximately equal to the total length of dysarthric examples used in the dysarthric-only condition. The training set is therefore composed of both dysarthric and typical utterances, but without increasing the total amount of training data. This variant is intended to test whether replacing part of the dysarthric-only training set with typical speech changes model performance when the total training-set size is controlled.

The second variant adds typical speech data directly to the available dysarthric training data. In this setup, the typical recordings are treated as additional training examples, increasing the total amount of training data. This variant evaluates the practical effect of augmenting the dysarthric training set with typical speech.

Both variants use the same phoneme recognition architecture and contrastive learning framework as the dysarthric-only experiments. The typical and dysarthric recordings are phonemized using the same vocabulary and preprocessing pipeline, allowing them to be used jointly during training. In the contrastive learning setup, phoneme-level embeddings can therefore be extracted from both typical and dysarthric utterances and included in the same triplet construction procedure.

4

Implementing and Evaluating the Bilingual Contrastive Framework

4.1. Datasets

In the experiments within the scope of this thesis, three different datasets are used: one containing dysarthric speech data in two different languages, and subsets of two different datasets containing typical speech data - one in Dutch and one in English.

4.1.1. Dysarthric speech data

The dysarthric speech data used for the experiments comes exclusively from the DysOne dataset [69]. The dataset contains speech of a single dysarthric individual, a 35-year-old male, whose dysarthria has been the result of brain damage from high-impact trauma. The speech has been classified by a speech language pathologist as severe to very severe level and flaccid type. This results in his overall speech intelligibility being noticeably reduced compared to typical speech.

The dataset is composed of speech data in two different languages: Dutch and English, with Dutch being the speaker's first language, and English second.

Speech data in both languages consists of read and spontaneous speech portions. The prompts for read speech in Dutch come from the Corpus Gesproken Nederlands (CGN; Spoken Dutch Corpus) [104], and the prompts for read speech in English come from the LibriSpeech dataset [105], and they include sentences of various lengths.

In addition to read speech data, the dataset contains read speech data with participant-prepared sentences and spontaneous speech utterances, consisting of answers to open-ended questions.

Both English and Dutch portions of the dataset have been split into training, validation and testing sets using a 70/10/20 split by number of utterances. The training, validation

and testing sets used for bilingual experiments have been composed by combining the datasets in each language respectively together.

Table 4.1 summarizes the final DysOne splits used in the experiments.

Table 4.1: Duration of the DysOne data used in the experiments, separated by language, data split, and speech type.

Language	Split	Read (h)	Spontaneous (h)	Total (h)
Dutch	Training	1.62	5.35	6.97
	Validation	0.23	0.10	0.33
	Testing	0.45	0.20	0.65
	Total	**2.30**	**5.65**	**7.95**
English	Training	2.23	8.00	10.23
	Validation	0.31	0.17	0.48
	Testing	0.63	0.36	1.00
	Total	**3.17**	**8.54**	**11.70**
Dutch + English		**5.47**	**14.18**	**19.66**

4.1.2. TYPICAL SPEECH DATA

For the experiments which examine the effect of incorporating typical speech data into the training as described in Section 3.8, the training set is modified.

The typical speech data in Dutch comes from Corpus Gesproken Nederlands [104] and is selected so that the transcript of the utterances used match the utterances chosen as read speech utterances for the DysOne corpus. The English samples come from the LibriSpeech dataset [105], and they are selected on the same basis.

For the experiments which replace a part of dysarthric speech training data with typical speech training data, the total amount of dysarthric speech utterances length from Table 4.1 is used as a threshold. Then, until this threshold is met, the dysarthric-typical utterance pairs matched by transcription are added to the modified training dataset. These pairs are sorted so that the longest dysarthric utterances that have a typical pair are chosen first. Both validation and testing sets remain unchanged. Table 4.2 shows the final splits for this variant.

Table 4.2: Dataset composition for the matched-size typical–dysarthric training scenario. Part of the dysarthric training data is replaced with typical speech, while the validation and test sets remain unchanged and contain only dysarthric speech.

Language	Dysarthric training (h)	Typical training (h)	Total training (h)	Validation (h)	Testing (h)
Dutch	5.26	1.66	6.92	0.33	0.65
English	7.79	2.27	10.06	0.48	1.00
Bilingual	13.05	3.93	16.98	0.81	1.65

The creation of the dataset for the variant with a full-size dysarthric training dataset with additional typical data is more straightforward. In this case, the dataset is simply augmented with typical data utterances matched with dysarthric read speech samples. The final splits for this scenario are depicted in Table 4.3

Table 4.3: Dataset composition for the additional typical-speech training scenario. The full dysarthric training set is retained and supplemented with typical speech, while the validation and test sets remain unchanged and contain only dysarthric speech.

Language	Dysarthric training (h)	Typical training (h)	Total training (h)	Validation (h)	Testing (h)
Dutch	6.97	1.66	8.63	0.33	0.65
English	10.23	2.27	12.50	0.48	1.00
Bilingual	17.20	3.93	21.13	0.81	1.65

4.2. DATA PREPROCESSING

4.2.1. DATA AUGMENTATION

To increase the acoustic variability of the limited training data without requiring additional recordings, speed perturbation was applied to the training utterances. Speed perturbation has previously been used successfully in dysarthric ASR and provides a simple way of exposing the model to variation in speaking rate [54–56]. For each training utterance, two additional versions were created: one with the speech slowed down to 0.9× the original speed and one with the speech sped up to 1.1× the original speed. These relatively moderate perturbations were chosen to introduce additional variation while preserving the linguistic content of the utterance.

The augmentation was applied offline before training. This avoided repeatedly generating the same transformed audio during training and ensured that all experimental configurations were trained on the same augmented versions of the data. The augmented utterances were included in the training set together with the original recordings.

4.2.2. PHONEMIZATION

The datasets used in the experiments contain orthographic transcriptions rather than phoneme-level labels. Since both the CTC output vocabulary and the contrastive learning framework operate at the phoneme level, these transcriptions first had to be converted into phoneme sequences.

The conversion of word-level transcriptions to phoneme sequences was done by using the open-source phonemizer *espeak-ng* [106]. For English samples, the English rule set was used, while for Dutch samples, the Dutch rule set was used.

4.2.3. PHONEME POST-PROCESSING

The phonemization process produced sequences of space-separated phoneme tokens, with individual tokens ranging from one to three characters in length. The raw output did not contain explicit word-boundary markers, meaning that information about word

separation would otherwise be lost. Preserving these boundaries was useful for maintaining the interpretability of the phoneme sequences and for later error analysis. Two approaches for representing word boundaries were therefore explored:

1. Converting phoneme symbols of length longer than 1 character to a different, one-character symbol, removing spaces in between phonemes within the same word, and treating spaces as a word boundary. This approach would have been convenient in a sense that it would allow the phoneme error rate to be interpreted as de facto character error rate. However, the analysis of the errors would have required reverse mapping, making the results a little more complicated to interpret.

2. Adding an additional word boundary symbol '|', without making changes to multiple-character phoneme tokens.

Because there were no noticeable differences for the results of the CTC-only fine-tuning experiments for either of these approaches, method 2 was chosen to make the results more easily interpretable.

A further post-processing step was applied to selected composite phonemizer outputs. Some of these outputs occurred relatively rarely and represented combinations of phonetic units that could instead be expressed using more common phoneme tokens. Since phoneme-level representations are used both as the CTC prediction target and when constructing contrastive pairs, treating these rare composite forms as independent phoneme classes would increase target sparsity and reduce the number of training examples available for the corresponding phoneme units. Selected composite outputs were therefore decomposed into sequences of existing phoneme tokens.

In case of English, rhoticized vowel sequences and syllabic consonant outputs, such as /ɚ/, /ʊɹ/, /ɪɹ/, /ɛɹ/, /ɔːɹ/, /ɑːɹ/, /əl/, and /ɳ/ were mapped to sequences such as /əɹ/, /ʊ ɹ/, /ɪ ɹ/, /ɛ ɹ/, /ɔːɹ/, /ɑːɹ/, /ə l/, and /ə n/, respectively. Similarly, longer composite outputs such as /aɪɚ/ were decomposed into /a ɪ ɚ/.

Similar normalization was applied to selected Dutch outputs: /ɲ/ was mapped to /n j/ and /yʊ/ was mapped to /y ʊ/.

According to section 3.4, besides phonemes that are represented by a different IPA symbol in Dutch and English yet equivalent to each other, English and Dutch also include phonemes that share the same symbol in the IPA alphabet, yet their pronunciation in each of the languages differs significantly. This is the case for two of the phonemes: /w/ and /v/. Since the contrastive sampling strategy uses same-symbol phonemes as positive pairs, leaving these phones with identical labels would risk creating cross-lingual positives that share very little acoustic properties with each other, thus leading to the model learning wrong relationships between the phonemes. In order to prevent that, the tokens of those two phonemes were changed: /w/ was changed to /EN_w/ or /NL_w/, depending on the language of the utterance, and similarly, /v/ was changed to /EN_v/ or /NL_v/.

4.2.4. VOCABULARY CONSTRUCTION

After phonemization and post-processing, a separate phoneme vocabulary was constructed for each training setting. The vocabulary was computed from the corresponding

phonemized training CSV file by collecting the unique space-separated phoneme tokens appearing in the training split.

For the monolingual experiments, the Dutch and English vocabularies were computed from the Dutch and English training files, respectively. For the bilingual experiments, the vocabulary was computed from the combined English-Dutch training file, so it corresponds to the union of phoneme labels observed in both languages after normalization and language-specific relabeling.

The word boundary symbol | was included as a regular output token when present in the training data. The CTC blank symbol was handled separately as part of the CTC output space.

4.3. BATCH CONSTRUCTION

Mini-batches were precomputed before training and stored as JSON files containing utterance IDs (see Section 3.6 for details). This was done to keep the batch composition fixed across comparable experiments and to ensure that contrastive learning conditions used the intended batch structure.

For monolingual experiments, separate batch files were created for Dutch and English. The corresponding training CSV file was first sorted by utterance duration in descending order. Utterances were then grouped sequentially into fixed-size batches of six utterances. This produces monolingual batches containing only Dutch or only English utterances, depending on the training condition. This implementation matches the default setup for SpeechBrain wav2vec CTC recipe [107].

For bilingual contrastive learning experiments, the equivalence-guided bilingual batches described in Section 3.6 were used. Each bilingual batch contained six utterances, with three English and three Dutch utterances. These batches were constructed to increase the likelihood that the cross-lingual positive phoneme pairs, especially manually defined different-symbol equivalences, were available within the same mini-batch.

The same bilingual batch files were also used for the bilingual CTC-only fine-tuning baseline. Although this baseline does not use the contrastive objective, using the same batch composition makes the bilingual fine-tuning and bilingual contrastive learning conditions more directly comparable. This ensures that differences between the bilingual CTC baseline and the bilingual contrastive learning variants are not caused by differences in mini-batch construction.

4.4. BASELINE TRAINING CONFIGURATION

The baseline models were trained using the CTC objective described in Section 3.1. Separate CTC-only baselines were trained for Dutch, English, and bilingual English-Dutch phoneme recognition. These baselines serve two purposes. First, they provide the main comparison point for evaluating the effect of contrastive learning. Second, the best baseline checkpoints are used to generate the forced alignments required for phoneme-level embedding extraction.

All baseline models use the same wav2vec-based architecture. The monolingual baselines are trained using only the corresponding language-specific training data, while the bilingual baseline is trained on the combined English and Dutch training data. For

the bilingual baseline, the same precomputed bilingual batches used in the bilingual contrastive learning experiments are used, so that the differences between CTC-only and contrastive learning experiments are not caused by different batch composition.

The baseline models were trained using the SpeechBrain wav2vec CTC recipe, with the hyperparameters kept at their default values. The pretrained acoustic encoder was initialized from `facebook/wav2vec2-xls-r-300m`, a multilingual wav2vec-based model. Since the experiments involve both English and Dutch speech, a multilingual pretrained encoder is better suited to a bilingual phoneme recognition setup than a monolingual acoustic model. The same baseline configuration was used across the monolingual and bilingual CTC-only experiments to ensure that differences in performance could be attributed to the training data and experimental condition rather than to changes in the model configuration.

The training configuration is summarized in Table 4.4.

Setting	Value
Pretrained encoder	`facebook/wav2vec2-xls-r-300m`
DNN layers	2
DNN hidden size	1024
Batch size	6
Optimizer	Adam
Optimizer weight decay	0.001
Learning rate for pretrained encoder	0.00005
Learning rate for task-specific layers	0.0003
Number of epochs	50
Checkpoint selection	Best validation PER

Table 4.4: Baseline training configuration.

4.5. PHONEME ALIGNMENT SETUP

Phoneme-level embeddings are extracted using CTC forced alignments, as described in Section 3.2. The alignments are generated before contrastive learning using the best checkpoint of the corresponding CTC baseline model. The resulting alignment files contain, for each utterance, the phoneme IDs, phoneme indices, start frames, end frames, language label, and alignment scores.

Two fixed span extraction variants are evaluated. The first uses non-blank spans from the CTC forced-alignment path. The second uses contiguous spans, where the end frame of each phoneme is defined as the start frame of the next phoneme. In both cases, the alignments are precomputed once and kept fixed during contrastive learning.

4.6. CONTRASTIVE LEARNING CONFIGURATION

The contrastive learning experiments extend the CTC baseline with the phoneme-level triplet loss described in Section 3.3. During training, the total objective combines the CTC

loss and the triplet loss. During validation and testing, only the CTC phoneme predictions are used for evaluation.

The triplet loss is computed on normalized phoneme-span embeddings extracted from the encoder output using the precomputed forced alignments.

The contrastive learning experiments vary across three main factors: the positive sampling strategy, the span extraction variant, and the triplet loss weight λ_{triplet}. The positive sampling strategies correspond to the basic same-symbol, equivalence-prioritized bilingual, and hierarchical bilingual strategies described in Section 3.5. The span extraction variants correspond to the non-blank and contiguous alignment variants described in Section 3.2. The triplet loss weight λ_{triplet} controls the contribution of the contrastive objective relative to the total training loss.

The contrastive learning experiments were evaluated across a range of triplet loss weights. For each contrastive learning condition, the value of λ_{triplet} was varied over the same set of values:

$$\lambda_{\text{triplet}} \in \{0.05, 0.1, 0.2, 0.3, 0.4\}.$$

The tested range was chosen to cover relatively weak to stronger contributions of the contrastive objective while keeping the CTC loss as the primary training objective. This range was used consistently across the tested experimental dimensions, including the different positive sampling strategies, span extraction variants, and weighted bilingual loss variants. Using the same set of λ_{triplet} values across conditions makes it possible to compare how sensitive each setup is to the weight of the contrastive objective.

Table 4.5 presents the configuration for the contrastive learning experiments in detail.

Setting	Value
Triplet distance	Cosine distance
Triplet margin m	0.2
Triplet loss weight λ_{triplet}	$\{0.05, 0.1, 0.2, 0.3, 0.4\}$
Embedding normalization	L2 normalization
Projection head	Not used
Contrastive learning schedule	Applied from the start of training
Alignment type	Fixed precomputed CTC alignments

Table 4.5: Contrastive learning configuration used in the main experiments.

4.7. WEIGHTED BILINGUAL CONTRASTIVE LOSS VARIANTS

In addition to the unweighted contrastive learning conditions, weighted bilingual contrastive loss variants were evaluated for the two bilingual sampling strategies introduced in Section 3.7.2. These variants use the same triplet construction procedure as their unweighted counterparts, but assign different weights to the triplet categories corresponding to the sampling priority levels.

For the equivalence-prioritized bilingual sampling strategy, two priority levels are used: manually defined cross-lingual different-symbol positives and same-symbol posi-

tives. The weighted equivalence-prioritized loss was evaluated using weights of 0.8 and 0.2, respectively. This gives higher importance to the explicitly bilingual equivalence-based triplets while still utilizing same-symbol positives.

For the hierarchical bilingual sampling strategy, three priority levels are used: manually defined cross-lingual different-symbol positives, cross-lingual same-symbol positives, and same-language same-symbol positives. Two weighting schemes were evaluated. The first uses weights of 0.6, 0.3, and 0.1, providing a gradual decrease across the three priority levels. The second uses weights of 0.8, 0.15, and 0.05, placing stronger emphasis on manually defined cross-lingual different-symbol positives.

4.8. EXPERIMENTAL CONDITIONS

The experiments were organized into two main groups. The first group used only dysarthric speech data, while the second group investigated the effect of incorporating typical speech data. Within each group, the experiments were designed to first establish CTC-only baselines and then evaluate bilingual contrastive learning variants.

The dysarthric-only experiments were conducted first. The initial experiments consisted of CTC-only fine-tuning for monolingual Dutch, monolingual English, and bilingual English-Dutch data. Including monolingual baselines alongside the bilingual ones makes it possible to separate two distinct sources of improvement: gains from phoneme-level contrastive learning itself, which may also appear in a monolingual setting, from gains associated with joint bilingual training and, in the contrastive conditions, the use of cross-lingual positive pairs. These experiments served as baselines for evaluating the effect of adding the phoneme-level contrastive objective. After this, basic contrastive learning experiments were performed for monolingual Dutch, monolingual English, and bilingual English-Dutch data. In the basic contrastive learning setup, positives were selected using same-symbol sampling only.

The next set of dysarthric-only experiments evaluated the bilingual positive sampling strategies. First, the unweighted equivalence-prioritized and hierarchical sampling strategies were tested. After evaluating the unweighted variants, weighted versions of both strategies were tested. The weighted equivalence-prioritized setup used two loss weights corresponding to its two priority levels, while the weighted hierarchical setup used three loss weights corresponding to its three priority levels, as described in Section 4.7.

After the dysarthric-only experiments, the same general experimental structure was applied to the two typical-speech conditions described in Section 3.8. For each typical-data setup, CTC-only fine-tuning, basic contrastive learning, equivalence-prioritized setup, and hierarchical setup were evaluated. However, for the last two, only the variants that performed best in the dysarthric-only experiments were retained. Specifically, the typical data experiments used the weighted equivalence-prioritized setup and the unweighted hierarchical setup.

Across all contrastive learning experiments, two fixed alignment variants were evaluated: non-blank CTC spans and contiguous phoneme spans. In addition, each contrastive learning condition was evaluated using the same set of triplet loss weights:

$$\lambda_{\text{triplet}} \in \{0.05, 0.1, 0.2, 0.3, 0.4\}.$$

This made it possible to compare the effect of positive sampling strategy, loss weighting,

alignment type, and contrastive loss strength across the experimental conditions. The overview of the experimental conditions is presented in Table 4.6.

Data setup	Experiment type	Conditions tested
Dysarthric-only	CTC-only baseline	Dutch, English, bilingual
Dysarthric-only	Basic contrastive learning	Dutch, English, bilingual; same-symbol positives
Dysarthric-only	Equivalence-prioritized bilingual sampling	Bilingual; unweighted and weighted variants
Dysarthric-only	Hierarchical bilingual sampling	Bilingual; unweighted and weighted variants
Matched-size typical–dysarthric	Typical-data experiments	CTC-only, basic CL, weighted equivalence-prioritized setup, unweighted hierarchical setup
Typical-data addition	Typical-data experiments	CTC-only, basic CL, weighted equivalence-prioritized setup, unweighted hierarchical setup

Table 4.6: Overview of experimental conditions evaluated in this thesis.

4.8.1. SOFTWARE AND HARDWARE INFRASTRUCTURE

All of the experiments were run on the Delft AI Cluster (DAIC), using an NVIDIA A40 GPU unit for training. All of the code was implemented in Python 3.9.23 using PyTorch 2.6.0. The code was built on top of SpeechBrain's wav2vec CTC recipe, with hyperparameters managed through SpeechBrain's YAML configuration system, and the wav2vec-XLS-R encoder loaded through the Hugging Face library.

4.9. EVALUATION METRICS

4.9.1. PHONEME ERROR RATE

The primary evaluation metric is phoneme error rate (PER). Since the recognition model is trained to predict phoneme sequences and the proposed contrastive objective operates directly on phoneme-level representations, PER provides a measure of recognition performance at the same linguistic level targeted by the proposed method. PER is computed as:

$$\mathrm{PER} = \frac{S + D + I}{N} \times 100,$$

where S, D, and I denote the number of substitutions, deletions, and insertions, respectively, and N is the number of reference phonemes.

PER is computed for several evaluation scopes. First, overall PER is computed over the complete test set. Second, language-specific PER is computed separately for English and Dutch test utterances. Third, equivalence-phoneme PER is computed only over the phonemes involved in the manually defined cross-lingual equivalence mapping. The

equivalence-phoneme subset is particularly important because these phonemes are directly targeted by the explicitly cross-lingual contrastive strategies. Evaluating them separately makes it possible to determine whether effects that may be small at the level of the complete test set are concentrated on the phoneme relationships that the method is designed to modify.

For equivalence-specific PER, substitutions and deletions are counted when the reference phoneme belongs to the target equivalence-phoneme set. Insertions are counted when the inserted hypothesis phoneme belongs to the target set. The same convention is used for individual phoneme-level error analyses.

In addition to PER, substitution, insertion, and deletion rates are reported separately. This makes it possible to analyze which type of recognition error contributes most strongly to the observed PER changes.

4.9.2. PHONEME-LEVEL REPRESENTATION ANALYSIS

PER measures whether changes to the training objective affect recognition performance, but it does not reveal whether contrastive learning changes the structure of the learned phoneme representations in the intended way. Representation space analysis is therefore used to examine the mechanism targeted directly by the contrastive objective. Three complementary analyses are applied to the extracted phoneme-level embeddings: cosine distance, nearest-neighbor analysis, and silhouette score.

Each of these metrics measures a different aspect of the embedding space – cosine distance measures pairwise similarity, nearest-neighbor analysis captures local neighborhood composition, and silhouette score provides a measure of cluster separation. These metrics do not always agree – for example, a category's average distance can shrink without that relationship dominating an embedding's local neighborhood, and local neighbor structure can shift without a corresponding change in overall cluster separation – therefore, using all three gives a more complete evaluation of how contrastive learning restructures phoneme representations at the pairwise, local, and global levels.

COSINE DISTANCE

Cosine-distance analysis is used to examine whether different categories of phoneme-level embedding pairs are closer or farther apart in the learned representation space.

Since the triplet objective employed in this thesis is itself defined using cosine distance, evaluating the representations with the same distance measure therefore allows the analysis to directly examine the relationships optimized during contrastive training.

For two embeddings $\mathbf{z}_i$ and $\mathbf{z}_j$, cosine distance is defined as:

$$d_{\cos}(\mathbf{z}_i, \mathbf{z}_j) = 1 - \mathbf{z}_i^\top \mathbf{z}_j.$$

Embedding pairs were assigned to predefined categories, including same-phoneme same-language pairs, cross-lingual same-symbol pairs, manually defined cross-lingual equivalent pairs, and unrelated pairs. For each category C, the average cosine distance was computed as:

$$\bar{d}_C = \frac{1}{|C|} \sum_{(i,j)\in C} d_{\cos}(\mathbf{z}_i, \mathbf{z}_j).$$

Lower values of $\bar{d}_C$ indicate that embeddings in that category are more similar. If the contrastive objective successfully shapes the representation space according to the defined positive relationships, positive-pair categories are expected to have lower average cosine distance than unrelated pairs.

NEAREST NEIGHBORS

Nearest-neighbor analysis is used to examine the local structure of the learned phoneme-level embedding space. For each embedding $\mathbf{z}_i$, the k closest embeddings are retrieved using cosine distance, excluding the anchor embedding itself. While average cosine distance describes the overall proximity of a pair category, nearest-neighbor analysis tests whether these relationships are sufficiently strong to determine the immediate local neighborhood of an embedding.

For each anchor i and pair category c, let $n_c(i)$ denote the number of its k nearest neighbors that belong to category c. The proportion of nearest neighbors belonging to that category is then computed as:

$$p_c(i) = \frac{n_c(i)}{k}.$$

The considered categories are same-phoneme same-language pairs, cross-lingual same-symbol pairs, and manually defined English–Dutch phoneme-equivalence pairs.

The proportions are then averaged across all M anchor embeddings:

$$\bar{p}_c = \frac{1}{M} \sum_{i=1}^{M} p_c(i).$$

Higher values of $\bar{p}_c$ indicate that phoneme pairs belonging to category c occur more frequently in the local neighborhoods of the embeddings. This provides a measure of whether the contrastive objective brings the targeted phoneme relationships closer together locally in the representation space.

SILHOUETTE SCORE

Silhouette score, introduced by Rousseeuw [108], is used to quantify how compact and well separated the phoneme-label groups are in the learned representation space. For each embedding, the score compares its average distance to embeddings with the same phoneme label with its average distance to the nearest alternative phoneme group. It therefore captures two properties of the representation space: cohesion, describing how closely embeddings with the same label are grouped together, and separation, describing how distinct that group is from other phoneme groups. The score ranges from -1 to 1, where values close to 1 indicate compact and well-separated groups, values around 0 indicate substantial overlap between groups, and negative values indicate that an embedding is, on average, closer to another phoneme group than to its own [109].

For each embedding i, the silhouette score compares the average distance to embeddings with the same phoneme label, $d_{\text{intra}}(i)$, with the lowest average distance to embeddings from another phoneme cluster, $d_{\text{inter}}(i)$):

$$\text{sil}(i) = \frac{d_{\text{inter}}(i) - d_{\text{intra}}(i)}{\max(d_{\text{intra}}(i), d_{\text{inter}}(i))}.$$

The final silhouette score is obtained by averaging sil(i) over all included embeddings. Values closer to 1 indicate well-separated phoneme clusters, whereas lower values indicate overlapping clusters, or that embeddings are closer to another phoneme cluster than to their assigned one.

In this thesis, three silhouette-based measures are reported. The general silhouette score uses exact phoneme labels as cluster labels and measures general phoneme-level separation. The equivalence-aware silhouette score merges manually defined English-Dutch equivalent phonemes into shared cluster labels, while keeping the rest of the phoneme labels unchanged, testing whether equivalent phonemes cluster better when treated as one category. The equivalence-only silhouette score is computed only on phonemes that belong to the manual equivalence mapping, serving as a measure of how well the cross-lingual equivalence groups are organized.

4.9.3. STATISTICAL SIGNIFICANCE TESTING

Since all systems are evaluated on the same test utterances, statistical significance is evaluated using paired non-parametric bootstrap resampling over test utterances, following general practice for ASR significance testing [110, 111]. The bootstrap samples are paired so that each resampled utterance is included for both systems. This ensures that differences in PER are due to differences between the systems rather than differences in which utterances were sampled.

At each bootstrap iteration, utterances are sampled with replacement from the aligned test set. Aggregate PER is recomputed by pooling the total number of phoneme errors and reference phonemes across the sampled utterances.

For each resample, the tested statistic is the PER difference between the two systems:

$$\Delta\text{PER} = \text{PER}_{\text{II}} - \text{PER}_{\text{I}}.$$

Negative values indicate that system II obtains lower PER than system I. For each comparison, 10,000 bootstrap samples are drawn using a fixed random seed. The 2.5th and 97.5th percentiles of the resulting bootstrap distribution define the 95% confidence interval. A difference is treated as statistically significant when this interval does not include zero. The same procedure is applied to both the main PER results and the equivalence-phoneme PER results.

5

Recognition and Representation Space Results

For each contrastive learning condition, multiple configurations were evaluated by varying the phoneme-span extraction method and the triplet-loss weight λ_{triplet}. The main results reported in this chapter show the configuration with the lowest observed test-set PER for each approach, while the complete results across all tested configurations are provided in Appendix A. Unless stated otherwise, references to the "best" configuration therefore refer to this selection procedure.

5.1. Results for dysarthric-data-only experiments

5.1.1. Phoneme error rate results

Table 5.1 compares the monolingual CTC-only baselines with the monolingual contrastive learning approaches. In both languages, contrastive learning obtains slightly lower observed overall PER than CTC-only fine-tuning. For Dutch, PER is 11.29% with contrastive learning compared with 11.44% for CTC-only fine-tuning. For English, the difference is larger, with PER of 8.23% and 8.55%, respectively. However, neither reduction is statistically significant (see Table 5.7).

Table 5.2 compares the bilingual CTC-only baseline with the bilingual basic same-symbol sampling strategy. Basic contrastive learning obtains lower observed PER than the CTC-only baseline across all evaluation scopes. Overall PER decreases from 10.46% to 9.96%, while Dutch and English PER decrease from 12.30% to 11.46% and from 8.88% to 8.65%, respectively. Lower PER is also observed for the equivalence-phoneme subsets. The reductions in overall PER and overall equivalence PER are statistically significant (see Table 5.7).

Table 5.3 compares the three bilingual contrastive learning strategies without the weighted loss applied: basic same-symbol sampling, equivalence-prioritized sampling, and hierarchical sampling. The basic strategy obtains the lowest observed overall PER at 9.96%, followed by hierarchical sampling with 10.02% and equivalence-prioritized

Table 5.1: Best monolingual CTC-only and monolingual contrastive learning results for dysarthric-data-only experiments. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. Equiv. PER denotes PER computed over the language-specific phoneme equivalence set. The best result in each comparison is underlined.

Language	Evaluation scope	CTC-only	Best basic CL
English	PER	**8.55** S: 4.76 I: 1.56 D: 2.24	**<u>8.23</u>** S: 4.65 I: 1.46 D: 2.13
English	Equiv. PER	**14.27** S: 10.19 I: 1.49 D: 2.58	**<u>12.36</u>** S: 9.10 I: 1.22 D: 2.04
Dutch	PER	**11.44** S: 5.78 I: 2.17 D: 3.49	**<u>11.29</u>** S: 5.83 I: 2.09 D: 3.37
Dutch	Equiv. PER	**<u>10.91</u>** S: 6.85 I: 1.42 D: 2.63	**11.85** S: 7.76 I: 1.90 D: 2.20

Table 5.2: Best dysarthric-data-only bilingual CTC-only and basic same-symbol contrastive learning results. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. Equiv. denotes evaluation over the phoneme equivalence set. The best result in each evaluation scope is underlined.

Evaluation scope	CTC-only	Best basic CL
Overall	**10.46** S: 5.74 I: 1.94 D: 2.78	**<u>9.96</u>** S: 5.60 I: 1.65 D: 2.71
Dutch	**12.30** S: 6.48 I: 2.35 D: 3.47	**<u>11.46</u>** S: 6.27 I: 1.88 D: 3.32
English	**8.88** S: 5.11 I: 1.59 D: 2.18	**<u>8.65</u>** S: 4.98 I: 1.43 D: 2.24
Equiv.	**12.81** S: 8.38 I: 1.85 D: 2.59	**<u>11.76</u>** S: 8.08 I: 1.43 D: 2.27
EN equiv.	**12.63** S: 8.46 I: 2.18 D: 1.98	**<u>11.67</u>** S: 7.92 I: 1.91 D: 1.84
NL equiv.	**12.93** S: 8.32 I: 1.64 D: 2.97	**<u>11.64</u>** S: 7.54 I: 1.16 D: 2.93

sampling with 10.15%. Neither of the explicitly bilingual strategies differs significantly from basic same-symbol sampling in overall PER, meaning that there is no evidence that explicitly prioritizing cross-lingual positive relationships improves overall recognition over basic same-symbol sampling.

For the equivalence-phoneme subset, hierarchical sampling obtains lower observed PER than basic same-symbol sampling (11.23% compared with 11.76%). However, this difference is not statistically significant (95% CI [-1.38, +0.33]; see Table 5.7).

Table 5.3: Best dysarthric-data-only bilingual contrastive learning results across sampling strategies. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. Equiv. denotes evaluation over the phoneme equivalence set. The best result in each evaluation scope is underlined.

Evaluation scope	Basic CL	Equiv.-prioritized CL	Hierarchical CL
Overall	**9.96** S: 5.60 I: 1.65 D: 2.71	**10.15** S: 5.64 I: 1.78 D: 2.72	**10.02** S: 5.48 I: 1.73 D: 2.80
Dutch	**11.46** S: 6.27 I: 1.88 D: 3.32	**11.69** S: 6.19 I: 2.11 D: 3.38	**11.47** S: 6.01 I: 1.92 D: 3.53
English	**8.65** S: 4.98 I: 1.43 D: 2.24	**8.76** S: 5.07 I: 1.36 D: 2.32	**8.64** S: 4.97 I: 1.43 D: 2.24
Equiv.	**11.76** S: 8.08 I: 1.43 D: 2.27	**11.76** S: 8.11 I: 1.51 D: 2.25	**11.23** S: 7.77 I: 1.16 D: 2.40
EN equiv.	**11.67** S: 7.92 I: 1.91 D: 1.84	**12.01** S: 8.67 I: 1.71 D: 1.64	**11.54** S: 7.92 I: 1.57 D: 2.05
NL equiv.	**11.64** S: 7.54 I: 1.16 D: 2.93	**11.38** S: 7.67 I: 1.21 D: 2.50	**11.21** S: 7.67 I: 0.91 D: 2.63

Table 5.4 compares the unweighted and weighted versions of the equivalence-prioritized strategy. The weighted version obtains numerically lower overall, Dutch, English, and English-equivalence PER, while the unweighted version obtains lower PER for the overall equivalence set and the Dutch equivalence subset. For the two statistically tested scopes, overall PER and overall equivalence PER, the differences between the weighted and unweighted configurations are not statistically significant.

Table 5.5 shows a different pattern for hierarchical bilingual sampling. The unweighted configuration obtains lower overall PER than both weighted variants. The difference in overall PER between the unweighted configuration and weighted Version 1 is not statistically significant, whereas weighted Version 2 produces a statistically significant increase in overall PER relative to the unweighted configuration.

Table 5.6 summarizes the results across the main training strategies. The equivalence-

Table 5.4: Best dysarthric-data-only equivalence-prioritized contrastive learning results for the unweighted and weighted objectives. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. Equiv. denotes evaluation over the phoneme equivalence set. The best result in each evaluation scope is underlined.

Evaluation scope	Unweighted equiv.-prioritized CL	Weighted equiv.-prioritized CL
Overall	**10.15** S: 5.64 I: 1.78 D: 2.72	**<u>9.94</u>** S: 5.48 I: 1.79 D: 2.66
Dutch	**11.69** S: 6.19 I: 2.11 D: 3.38	**<u>11.39</u>** S: 5.98 I: 2.13 D: 3.27
English	**8.76** S: 5.07 I: 1.36 D: 2.32	**<u>8.52</u>** S: 4.92 I: 1.39 D: 2.21
Equiv.	**<u>11.76</u>** S: 8.11 I: 1.51 D: 2.25	**11.92** S: 7.98 I: 1.51 D: 2.40
EN equiv.	**12.01** S: 8.67 I: 1.71 D: 1.64	**<u>11.67</u>** S: 7.99 I: 1.57 D: 2.12
NL equiv.	**<u>11.38</u>** S: 7.67 I: 1.21 D: 2.50	**11.51** S: 7.59 I: 1.34 D: 2.59

Table 5.5: Best dysarthric-data-only hierarchical contrastive learning results for the unweighted and weighted objectives. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. Equiv. denotes evaluation over the phoneme equivalence set. The best available result in each evaluation scope is underlined.

Evaluation scope	Unweighted hierarchical CL	Weighted hierarchical CL ver. 1	Weighted hierarchical CL ver. 2
Overall	**<u>10.02</u>** S: 5.48 I: 1.73 D: 2.80	**10.24** S: 5.56 I: 1.85 D: 2.83	**10.27** S: 5.65 I: 1.86 D: 2.75
Dutch	**<u>11.47</u>** S: 6.01 I: 1.92 D: 3.53	**11.68** S: 6.04 I: 2.14 D: 3.49	**11.60** S: 5.93 I: 2.24 D: 3.43
English	**<u>8.64</u>** S: 4.97 I: 1.43 D: 2.24	**9.02** S: 5.22 I: 1.53 D: 2.26	**8.97** S: 5.09 I: 1.49 D: 2.39
Equiv.	**<u>11.23</u>** S: 7.77 I: 1.16 D: 2.40	**11.92** S: 8.16 I: 1.64 D: 2.14	**11.73** S: 8.30 I: 1.51 D: 1.93
EN equiv.	**11.54** S: 7.92 I: 1.57 D: 2.05	**11.74** S: 8.19 I: 1.37 D: 2.18	**<u>11.47</u>** S: 7.99 I: 1.77 D: 1.71
NL equiv.	**<u>11.21</u>** S: 7.67 I: 0.91 D: 2.63	**11.55** S: 7.80 I: 1.25 D: 2.50	**11.29** S: 7.59 I: 1.38 D: 2.33

prioritized strategy obtains the lowest observed overall PER at 9.94%, followed by basic contrastive learning at 9.96% and hierarchical contrastive learning at 10.02%. However, the differences between the three strategies are not statistically significant. Similarly, no statistically significant differences are observed between the tested contrastive learning strategies for the equivalence-phoneme subset. Overall, the results do not provide evidence that one of the bilingual contrastive learning strategies consistently outperforms the others.

Table 5.6: Best dysarthric-data-only bilingual results across the main training strategies. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. Equiv. denotes evaluation over the phoneme equivalence set. The best result in each evaluation scope is underlined.

Scope	CTC-only bilingual	Basic CL	Equivalence prioritized CL	Hierarchical CL
Overall	**10.46** S: 5.74 I: 1.94 D: 2.78	**9.96** S: 5.60 I: 1.65 D: 2.71	**9.94** S: 5.48 I: 1.79 D: 2.66	**10.02** S: 5.48 I: 1.73 D: 2.80
Dutch	**12.30** S: 6.48 I: 2.35 D: 3.47	**11.46** S: 6.27 I: 1.88 D: 3.32	**11.39** S: 5.98 I: 2.13 D: 3.27	**11.47** S: 6.01 I: 1.92 D: 3.53
English	**8.88** S: 5.11 I: 1.59 D: 2.18	**8.65** S: 4.98 I: 1.43 D: 2.24	**8.52** S: 4.92 I: 1.39 D: 2.21	**8.64** S: 4.97 I: 1.43 D: 2.24
Equiv.	**12.81** S: 8.38 I: 1.85 D: 2.59	**11.76** S: 8.08 I: 1.43 D: 2.27	**11.76** S: 8.11 I: 1.51 D: 2.25	**11.23** S: 7.77 I: 1.16 D: 2.40
EN equiv.	**12.63** S: 8.46 I: 2.18 D: 1.98	**11.67** S: 7.92 I: 1.91 D: 1.84	**11.67** S: 7.99 I: 1.57 D: 2.12	**11.47** S: 7.99 I: 1.77 D: 1.71
NL equiv.	**12.93** S: 8.32 I: 1.64 D: 2.97	**11.64** S: 7.54 I: 1.16 D: 2.93	**11.38** S: 7.67 I: 1.21 D: 2.50	**11.21** S: 7.67 I: 0.91 D: 2.63

It is worth noting that the best bilingual systems do not outperform the best monolingual contrastive learning systems when Dutch and English are evaluated separately.

Full phoneme error rate results for experiments using dysarthric data only can be found in Section A.1 of the Appendix.

Table 5.7: Paired bootstrap significance testing for selected PER comparisons for dysarthric-data-only scenario experiments. ΔPER is defined as $PER_{II} - PER_{I}$; negative values indicate that System II performs better. A difference is considered statistically significant when the 95% confidence interval does not include zero.

Scope	System I	System II	PER I	PER II	ΔPER	95% CI	Sig.
Overall	CTC-only monolingual Dutch	Basic CL monolingual Dutch	11.44	11.29	-0.15	[-0.64, +0.33]	**No**
Overall	CTC-only monolingual English	Basic CL monolingual English	8.55	8.23	-0.32	[-0.72, +0.08]	**No**
Overall	CTC-only	Basic CL	10.46	9.96	-0.50	[-0.81, -0.18]	**Yes**
Equiv.	CTC-only	Basic CL	12.81	11.76	-1.06	[-1.88, -0.22]	**Yes**
Overall	Basic CL	Equiv.-prioritized CL best setup	9.96	9.94	-0.02	[-0.35, +0.30]	**No**
Equiv.	Basic CL	Equiv.-prioritized CL best setup	11.76	11.76	0.00	[-0.82, +0.80]	**No**
Overall	Basic CL	Hierarchical CL best setup	9.96	10.02	+0.06	[-0.26, +0.38]	**No**
Equiv.	Basic CL	Hierarchical CL best setup	11.76	11.23	-0.53	[-1.38, +0.33]	**No**
Overall	Equiv.-prioritized CL unweighted	Equiv.-prioritized CL weighted	10.15	9.94	-0.21	[-0.54, +0.12]	**No**
Equiv.	Equiv.-prioritized CL unweighted	Equiv.-prioritized CL weighted	11.76	11.92	+0.16	[-0.70, +0.99]	**No**
Overall	Hierarchical CL unweighted	Hierarchical CL weighted ver. 1	10.02	10.24	+0.23	[-0.09, +0.54]	**No**
Equiv.	Hierarchical CL unweighted	Hierarchical CL weighted ver. 1	11.23	11.92	+0.69	[-0.20, +1.60]	**No**
Overall	Hierarchical CL unweighted	Hierarchical CL weighted ver. 2	10.02	10.27	+0.45	[+0.11, +0.78]	**Yes**
Equiv.	Hierarchical CL unweighted	Hierarchical CL weighted ver. 2	11.23	11.73	+0.74	[-0.11, +1.61]	**No**
Overall	Equiv.-prioritized CL best setup	Hierarchical CL best setup	9.94	10.02	+0.08	[-0.24, +0.41]	**No**
Equiv.	Equiv.-prioritized CL best setup	Hierarchical CL best setup	11.76	11.23	-0.53	[-1.29, +0.24]	**No**

5.1.2. EMBEDDING SPACE ANALYSIS

Table 5.8: Mean cosine distance between phoneme-span embeddings for different pair categories across systems for dysarthric-data-only scenario experiments. Lower values indicate that the corresponding pair type is represented more closely in the embedding space.

System	Same phoneme same language	Cross-lingual same symbol	Cross-lingual equivalences	Unrelated
Bilingual CTC	0.241	0.427	0.657	0.737
Basic CL	0.196	0.350	0.618	0.714
Equivalence prioritized unweighted	0.215	0.383	0.633	0.748
Equivalence prioritized weighted	0.197	0.327	**0.451**	0.724
Hierarchical unweighted	0.203	0.365	0.647	0.758
Hierarchical weighted ver. 1	**0.188**	**0.270**	0.463	**0.781**
Hierarchical weighted ver. 2	0.209	0.306	0.475	0.770

Table 5.9: Silhouette scores for phoneme-span embeddings across systems for dysarthric-data-only scenario experiments. The score was computed using phoneme labels as cluster labels and cosine distance as the distance metric. Higher values indicate more compact and better-separated phoneme-level clusters. CTC denotes bilingual CTC-only fine-tuning, Basic denotes basic same-symbol contrastive learning, A denotes the equivalence-prioritized strategy, and B denotes the hierarchical strategy.

System	Silhouette score	Equivalence-aware silhouette score	Equivalence-only silhouette score
Bilingual CTC	0.330	0.333	0.354
Basic CL	0.403	0.403	0.397
Equivalence prioritized unweighted	0.379	0.377	0.352
Equivalence prioritized weighted	0.441	0.433	0.459
Hierarchical unweighted	0.407	0.404	0.380
Hierarchical weighted ver. 1	**0.452**	**0.448**	**0.490**
Hierarchical weighted ver. 2	0.412	0.410	0.463

The embedding-space analysis provides additional insight into how the different training objectives are associated with the learned phoneme representations. Table 5.8 shows that the ordering of pair categories is the same across systems: same-phoneme same-language pairs have the lowest cosine distance, unrelated pairs have the highest distance, and cross-lingual pair categories fall between them. Compared with the bilingual CTC-only baseline, all contrastive learning approaches obtain lower mean distances between same-phoneme embeddings (column 1). This pattern is consistent with same-phoneme representations being more closely grouped under contrastive learning.

The cosine-distance results also show lower average distances for cross-lingual equivalent phonemes in the weighted bilingual variants, both for same-symbol pairs and manually defined equivalence pairs. In particular, the weighted equivalence-prioritized and weighted hierarchical systems obtain substantially lower observed distances for the cross-lingual equivalence category than the CTC-only baseline, the basic contrastive

Table 5.10: Nearest-neighbor proportions for phoneme-span embeddings at $k = 10$ for dysarthric-data-only scenario experiments. Values indicate the average proportion of nearest neighbors belonging to each category. Same phoneme denotes neighbors with the same phoneme label, cross-lingual same symbol denotes eligible cross-lingual same-symbol pairs, and cross-lingual equivalences denotes eligible manually defined cross-lingual equivalent pairs. Higher values indicate that the corresponding category appears more frequently in the local neighborhood of each embedding. CTC denotes bilingual CTC-only fine-tuning, Basic denotes basic same-symbol contrastive learning, A denotes the equivalence-prioritized strategy, and B denotes the hierarchical strategy.

System	Same phoneme	Cross-lingual same symbol	Cross-lingual equivalences
Bilingual CTC	0.885	0.009	0.003
Basic CL	0.890	0.010	0.004
Equivalence prioritized unweighted	0.887	0.011	0.003
Equivalence prioritized weighted	**0.891**	0.010	**0.005**
Hierarchical unweighted	0.889	0.011	0.003
Hierarchical weighted ver. 1	0.890	**0.017**	0.003
Hierarchical weighted ver. 2	0.887	**0.017**	**0.005**

learning system, and the corresponding unweighted bilingual variants. This pattern is consistent with the weighted objective encouraging the targeted cross-lingual phoneme representations to become closer. However, cross-lingual equivalent pairs remain further apart than same-phoneme same-language pairs, indicating that the representation space remains more strongly organized by exact phoneme identity than by manually defined cross-lingual equivalence.

Table 5.9 shows a similar pattern. All contrastive learning systems obtain higher observed silhouette scores than the bilingual CTC-only baseline, indicating more compact and separated phoneme-label groups according to this metric. The weighted contrastive learning systems obtain the highest silhouette scores among the evaluated models.

The equivalence-aware silhouette scores are lower than the corresponding standard silhouette scores, indicating that merging manually defined cross-lingual equivalent phonemes into shared groups does not produce more clearly separated groups according to this metric.

The equivalence-only silhouette scores are highest for the weighted bilingual contrastive learning approaches. Together with the lower cosine distances observed for cross-lingual equivalence pairs, this pattern is consistent with stronger grouping of the targeted cross-lingual phonemes in these systems.

Table 5.10 shows that the local neighborhoods are dominated by same-phoneme representations. Cross-lingual same-symbol and manually defined cross-lingual equivalent phonemes only rarely appear among the 10 nearest neighbors. The weighted approaches obtain higher observed proportions of cross-lingual equivalent neighbors, while the weighted hierarchical approaches also obtain higher proportions of cross-lingual same-symbol neighbors. However, the absolute proportions remain low.

5.2. RESULTS FOR EXPERIMENTS INCLUDING ADDITIONAL TYPICAL DATA

This section presents the results of the experiments with additional typical speech data used during training.

5.2.1. MATCHED-SIZE TYPICAL-DYSARTHRIC SCENARIO

PHONEME ERROR RATE RESULTS

The first setup involves the matched-size typical-dysarthric scenario, where the total amount of training data is kept fixed and part of the dysarthric training data is replaced with typical speech data, testing whether typical speech can substitute for dysarthric speech when the amount of training data is controlled. In this setup, only the best equivalence-prioritized and hierarchical contrastive learning systems were evaluated – the analysis on the impact of weighted loss was not conducted for this setup.

Table 5.11 shows the monolingual results for the matched-size typical-dysarthric scenario. Compared with the dysarthric-data-only monolingual experiments, performance decreases substantially for both languages. For English, the best monolingual PER increases from 8.23% in the dysarthric-data-only setting to 10.67% in the matched-size typical-dysarthric setting. For Dutch, the best monolingual PER increases from 11.29% to 15.12%. Similar degradation can be observed for the equivalence-phoneme subsets. The deterioration relative to the dysarthric-data-only condition is statistically significant for both overall PER and equivalence-phoneme PER (see Table 5.13).

Table 5.11: Best monolingual CTC-only and monolingual contrastive learning results for matched-size typical-dysarthric data scenario experiments. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. Equiv. PER denotes PER computed over the language-specific phoneme equivalence set. The best result in each comparison is underlined.

Language	Evaluation scope	CTC-only	Best basic CL
English	PER	**10.67** S: 6.39 I: 2.03 D: 2.25	**10.92** S: 6.61 I: 1.97 D: 2.33
English	Equiv. PER	**15.35** S: 12.26 I: 1.77 D: 1.22	**14.95** S: 12.13 I: 1.36 D: 1.36
Dutch	PER	**15.22** S: 8.08 I: 2.66 D: 4.48	**15.12** S: 8.27 I: 2.70 D: 4.15
Dutch	Equiv. PER	**15.22** S: 9.91 I: 1.55 D: 3.75	**15.65** S: 10.22 I: 1.85 D: 3.58

Table 5.12 shows the bilingual results for the matched-size typical-dysarthric scenario. The same overall pattern is observed as in the monolingual experiments: performance

is considerably worse than in the dysarthric-data-only setting. The best overall bilingual PER is 13.08%, compared with 9.94% in the dysarthric-data-only experiments. The equivalence-phoneme results show the same trend, with the best overall equivalence PER increasing from 11.23% to 15.88%.

Within the matched-size bilingual setting, the differences between contrastive learning strategies are relatively small. The equivalence-prioritized strategy obtains the lowest overall PER, while the lowest values for the language-specific and equivalence-based evaluation scopes are distributed across the three contrastive learning strategies.

Table 5.12: Best bilingual results across the main training strategies for the matched-size typical-dysarthric data scenario. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. Equiv. denotes evaluation over the phoneme equivalence set. The best result in each evaluation scope is underlined.

Scope	CTC-only bilingual	Basic CL	Equivalence prioritized CL	Hierarchical CL
Overall	**13.21** S: 7.66 I: 2.09 D: 3.46	**13.20** S: 7.68 I: 2.13 D: 3.39	**13.08** S: 7.66 I: 2.12 D: 3.30	**13.19** S: 7.72 I: 2.09 D: 3.38
Dutch	**15.06** S: 8.14 I: 2.35 D: 4.57	**14.95** S: 8.28 I: 2.23 D: 4.44	**14.88** S: 8.17 I: 2.35 D: 4.36	**14.84** S: 8.07 I: 2.35 D: 4.42
English	**11.63** S: 7.24 I: 1.87 D: 2.52	**11.14** S: 6.93 I: 1.84 D: 2.37	**11.36** S: 6.99 I: 1.95 D: 2.41	**11.46** S: 7.31 I: 1.85 D: 2.30
Equiv.	**16.88** S: 11.62 I: 2.11 D: 3.14	**16.12** S: 10.89 I: 2.01 D: 3.22	**15.88** S: 11.07 I: 1.77 D: 3.14	**16.04** S: 10.99 I: 2.03 D: 3.01
EN equiv.	**18.16** S: 12.76 I: 2.94 D: 2.46	**16.31** S: 11.95 I: 2.25 D: 2.12	**16.18** S: 12.01 I: 1.84 D: 2.32	**16.45** S: 12.15 I: 2.32 D: 1.98
NL equiv.	**16.08** S: 10.91 I: 1.59 D: 3.58	**15.09** S: 9.96 I: 1.64 D: 3.49	**15.26** S: 9.96 I: 1.77 D: 3.53	**15.43** S: 10.30 I: 1.59 D: 3.53

Full phoneme error rate results for experiments using matched-size typical-dysarthric data setup can be found in Section A.2.1 of the Appendix.

Table 5.13: Paired bootstrap significance testing for selected PER comparisons for matched-size typical-dysarthric data scenario experiments. ΔPER is defined as $PER_{II} - PER_{I}$; negative values indicate that System II performs better. A difference is considered statistically significant when the 95% confidence interval does not include zero.

Scope	System I	System II	PER I	PER II	ΔPER	95% CI	Sig.
Overall	Best Dutch dysarthric-only	Best Dutch matched-size	11.29	15.12	+3.83	[+2.75, +4.94]	**Yes**
Overall	Best English dysarthric-only	Best English matched-size	8.23	10.67	+2.44	[+1.70, +3.23]	**Yes**
Overall	Best bilingual dysarthric-only	Best bilingual matched-size	9.94	13.08	+3.12	[+2.42, +3.85]	**Yes**
Equiv.	Best bilingual dysarthric-only	Best bilingual matched-size	11.23	15.88	+4.65	[+3.40, +5.97]	**Yes**

5

EMBEDDING SPACE ANALYSIS

The embedding-space analyses for the matched-size typical-dysarthric scenario show some interesting properties. Table 5.14 shows that the previously observed structure of the embedding space differs here – unrelated phonemes still are the furthest apart, but it is the cross-lingual phonemes that are the closest. Even in the CTC-only training, this pattern presents itself, which suggests that it is not an effect of the contrastive objective.

The hierarchical contrastive learning system further reduces the distance for cross-lingual same-symbol pairs and manually defined cross-lingual equivalence pairs. However, this representation-level structure does not lead to improved dysarthric ASR performance. In fact, the matched-size typical-dysarthric systems perform substantially worse than the corresponding dysarthric-data-only systems in terms of PER.

Table 5.14: Mean cosine distance between phoneme-span embeddings for different pair categories across systems for matched-size typical-dysarthric scenario experiments. Lower values indicate that the corresponding pair type is represented more closely in the embedding space.

System	Same phoneme same language	Cross-lingual same symbol	Cross-lingual equivalences	Unrelated
Bilingual CTC	0.607	0.357	**0.230**	0.713
Basic CL	0.629	0.363	0.235	0.720
Equivalence prioritized	**0.433**	0.308	0.231	0.755
Hierarchical	0.447	0.230	**0.201**	**0.805**

The silhouette scores, however, show a very similar pattern as was observed before. The results in table 5.15 show that, just as before, contrastive learning systems obtain higher phoneme-label silhouette scores than the bilingual CTC-only system. The hierarchical strategy obtains the highest silhouette score. In this setup, equivalence-only silhouette scores are also higher for the equivalence-prioritized and hierarchical strategies than for CTC-only and basic contrastive learning. Also, for this setup as well, the equivalence-aware silhouette scores are lower than the corresponding standard phoneme-

label silhouette scores.

Table 5.15: Silhouette scores for phoneme-span embeddings across systems for matched-size typical-dysarthric data scenario experiments. The score was computed using phoneme labels as cluster labels and cosine distance as the distance metric. Higher values indicate more compact and better-separated phoneme-level clusters.

System	Silhouette score	Equivalence-aware silhouette score	Equivalence-only silhouette score
Bilingual CTC	0.349	0.342	0.330
Basic CL	0.355	0.348	0.333
Equivalence prioritized	0.381	0.374	0.440
Hierarchical	**0.456**	**0.423**	**0.453**

Table 5.16 shows that also the nearest-neighbor neighborhoods are shaped similarly as they were in the dysarthric-data-only setup. Approximately 85% of the ten nearest neighbors share the same phoneme label as the anchor. Cross-lingual same-symbol and manually defined cross-lingual equivalent neighbors appear more frequently than in the dysarthric-data-only setting, but their absolute proportions still remain low.

Table 5.16: Nearest-neighbor proportions for phoneme-span embeddings at $k = 10$. Values indicate the average proportion of nearest neighbors belonging to each category for matched-size dysarthric-typical data scenario experiments. Same phoneme denotes neighbors with the same phoneme label, cross-lingual same symbol denotes eligible cross-lingual same-symbol pairs, and cross-lingual equivalences denotes eligible manually defined cross-lingual equivalent pairs. Higher values indicate that the corresponding category appears more frequently in the local neighborhood of each embedding.

System	Same phoneme	Cross-lingual same symbol	Cross-lingual equivalences
Bilingual CTC	0.855	0.027	0.009
Basic CL	0.853	0.032	0.008
Equivalence prioritized	**0.856**	0.034	**0.010**
Hierarchical	0.852	**0.044**	0.009

5.2.2. FULL DYSARTHRIC DATASET WITH ADDITIONAL TYPICAL DATA SCENARIO

The second typical-data setup adds typical speech on top of the full dysarthric training set. This setting evaluates whether typical speech can be useful as an additional training material.

PHONEME ERROR RATE RESULTS

Table 5.17 shows the monolingual results for the typical-data addition scenario. Compared with the dysarthric-data-only monolingual experiments, performance improves for both languages. For English, the best overall PER decreases from 8.23% to 7.58%. For Dutch, the best overall PER decreases from 11.29% to 9.77%.

Unlike previous monolingual contrastive learning setups, in this case contrastive learning has a mixed effect. For English, it is the CTC-only system that obtains the best overall PER, unlike what was observed before. For Dutch, just as before, contrastive learning improves overall PER.

Table 5.17: Best monolingual CTC-only and monolingual contrastive learning results for additional typical data scenario experiments. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. Equiv. PER denotes PER computed over the language-specific phoneme equivalence set. The best result in each comparison is underlined.

Language	Evaluation scope	CTC-only	Best basic CL
English	PER	**7.58** S: 4.08 I: 1.53 D: 1.97	**7.68** S: 4.18 I: 1.51 D: 1.99
English	Equiv. PER	**11.96** S: 8.70 I: 1.49 D: 1.77	**10.60** S: 7.74 I: 1.36 D: 1.49
Dutch	PER	**10.14** S: 5.20 I: 1.77 D: 3.17	**9.77** S: 5.10 I: 1.85 D: 2.82
Dutch	Equiv. PER	**10.22** S: 6.38 I: 1.29 D: 2.54	**9.87** S: 6.12 I: 1.42 D: 2.33

Table 5.18 shows the bilingual results for the typical-data addition scenario. This setup gives the best overall performance among all of the evaluated bilingual training conditions. The best overall PER is 8.57%, compared with 9.94% in the dysarthric-data-only bilingual experiments. The same trend is observed for the equivalence-phoneme subsets, where the best overall equivalence PER decreases from 11.23% to 10.09%. Compared with the best dysarthric-data-only system, the reductions in both overall PER and equivalence-phoneme PER are statistically significant (see Table 5.19).

Unlike in the dysarthric-data-only experiments, the observed differences between the CTC-only and contrastive learning systems are small in this setting. For overall PER, the CTC-only system obtains 8.67%, compared with 8.70% for basic contrastive learning, 8.57% for equivalence-prioritized contrastive learning, and 8.61% for hierarchical contrastive learning. The equivalence-prioritized strategy obtains the lowest observed overall and Dutch PER, while the hierarchical strategy obtains the lowest observed English PER and the lowest PER across the equivalence-based evaluation scopes.

Full phoneme error rate results for experiments using additional typical data setup can be found in Section A.2.2 of the Appendix.

Table 5.18: Best bilingual results for additional typical data scenario across the main training strategies. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. Equiv. denotes evaluation over the phoneme equivalence set. The best result in each evaluation scope is underlined.

Scope	CTC-only bilingual	Basic CL	Equivalence prioritized CL	Hierarchical CL
Overall	**8.67** S: 4.90 I: 1.47 D: 2.30	**8.70** S: 4.89 I: 1.59 D: 2.23	**8.57** S: 4.90 I: 1.41 D: 2.26	**8.61** S: 4.82 I: 1.50 D: 2.28
Dutch	**10.30** S: 5.49 I: 1.67 D: 3.13	**10.18** S: 5.51 I: 1.71 D: 2.97	**9.92** S: 5.50 I: 1.55 D: 2.88	**9.94** S: 5.23 I: 1.75 D: 2.97
English	**7.29** S: 4.39 I: 1.30 D: 1.60	**7.37** S: 4.31 I: 1.28 D: 1.78	**7.29** S: 4.28 I: 1.32 D: 1.70	**7.22** S: 4.32 I: 1.33 D: 1.56
Equiv.	**10.81** S: 7.13 I: 1.32 D: 2.35	**10.44** S: 6.95 I: 1.37 D: 2.11	**10.30** S: 7.13 I: 1.27 D: 1.90	**10.09** S: 6.82 I: 1.40 D: 1.88
EN equiv.	**10.03** S: 7.17 I: 1.30 D: 1.57	**9.69** S: 6.62 I: 1.57 D: 1.50	**9.62** S: 6.62 I: 1.43 D: 1.57	**9.22** S: 6.14 I: 1.64 D: 1.43
NL equiv.	**11.29** S: 7.11 I: 1.34 D: 2.84	**10.86** S: 7.72 I: 1.08 D: 2.07	**10.13** S: 6.72 I: 1.08 D: 2.33	**9.49** S: 6.76 I: 1.43 D: 1.30

Table 5.19: Paired bootstrap significance testing for selected PER comparisons for additional typical data scenario experiments. ΔPER is defined as $PER_{II} - PER_{I}$; negative values indicate that System II performs better. A difference is considered statistically significant when the 95% confidence interval does not include zero.

Scope	System I	System II	PER I	PER II	ΔPER	95% CI	Sig.
Overall	Best Dutch dysarthric-only	Best Dutch additional-data	11.29	9.77	-1.52	[-2.12, -0.92]	**Yes**
Overall	Best English dysarthric-only	Best English additional-data	8.23	7.58	-0.65	[-1.10, -0.21]	**Yes**
Overall	Best bilingual dysarthric-only	Best bilingual additional-data	9.94	8.57	-1.36	[-1.71, -1.01]	**Yes**
Equiv.	Best bilingual dysarthric-only	Best bilingual additional-data	11.23	10.09	-1.16	[-2.12, -0.24]	**Yes**

EMBEDDING SPACE ANALYSIS

The embedding-space results for the typical-data addition scenario show a pattern similar to the matched-size typical-dysarthric scenario. Table 5.20 shows that cross-lingual same-symbol and manually defined cross-lingual equivalence pairs have lower mean cosine distance than same-phoneme same-language pairs, which, again, differs from the dysarthric-data-only setting, where same-phoneme same-language pairs were the closest category.

Table 5.20: Mean cosine distance between phoneme-span embeddings for different pair categories across systems for additional typical data experiments. Lower values indicate that the corresponding pair type is represented more closely in the embedding space.

System	Same phoneme same language	Cross-lingual same symbol	Cross-lingual equivalences	Unrelated
Bilingual CTC	0.688	0.401	0.249	**0.759**
Basic CL	0.695	0.406	0.233	0.755
Equivalence prioritized	0.509	0.243	0.178	0.555
Hierarchical	**0.209**	**0.098**	**0.078**	0.243

Table 5.21 shows that the equivalence-prioritized and hierarchical systems obtain higher phoneme-label silhouette scores than the CTC-only and basic contrastive learning systems. These higher observed scores are consistent with greater phoneme-level group separation in the explicitly bilingual contrastive learning systems. The equivalence-prioritized system obtains the highest standard silhouette score and the highest equivalence-only silhouette score.

However, the equivalence-aware silhouette scores are, as before, lower than the corresponding standard silhouette scores. This suggests that the cross-lingual equivalence structure is not reflected in improved global group separation according to the silhouette metric.

Table 5.21: Silhouette scores for phoneme-span embeddings across systems for additional typical data experiments. The score was computed using phoneme labels as cluster labels and cosine distance as the distance metric. Higher values indicate more compact and better-separated phoneme-level clusters.

System	Silhouette score	Equivalence-aware silhouette score	Equivalence-only silhouette score
Bilingual CTC	0.396	0.389	0.362
Basic CL	0.392	0.388	0.361
Equivalence prioritized	**0.435**	**0.421**	**0.396**
Hierarchical	0.423	0.400	0.368

Table 5.22 shows that the nearest-neighbor neighborhoods are still dominated by same-phoneme representations. The hierarchical system obtains a higher observed proportion of cross-lingual same-symbol neighbors than the other systems, although the absolute proportion remains low.

In this setup, the observed proportions of manually defined cross-lingual equivalent neighbors are lower than in the other data conditions. Although the hierarchical system again obtains the highest proportion of cross-lingual same-symbol neighbors, this proportion is lower than in the matched-size typical-data scenario. Together with the silhouette-score results, this suggests that the global group-separation measures and the local nearest-neighbor structure capture different aspects of the representation space.

Table 5.22: Nearest-neighbor proportions for phoneme-span embeddings at $k = 10$. Values indicate the average proportion of nearest neighbors belonging to each category for additional typical data experiments. Same phoneme denotes neighbors with the same phoneme label, cross-lingual same symbol denotes eligible cross-lingual same-symbol pairs, and cross-lingual equivalences denotes eligible manually defined cross-lingual equivalent pairs. Higher values indicate that the corresponding category appears more frequently in the local neighborhood of each embedding.

System	Same phoneme	Cross-lingual same symbol	Cross-lingual equivalences
Bilingual CTC	0.900	0.007	**0.003**
Basic CL	0.901	0.007	**0.003**
Equivalence prioritized	0.903	0.019	**0.003**
Hierarchical	**0.906**	**0.028**	0.002

6

Discussion

6.1. RQ: How does bilingual phoneme-level contrastive learning with cross-lingual positive pairs affect personalized dysarthric ASR and the learned representation space?

While bilingual phoneme-level CL shows general improvement in PER compared to bilingual CTC-only fine-tuning, cross-lingual positive pairs do not seem to be the driving force behind this improvement. The improvements between the best-performing model that implements cross-lingual positive pairs and the bilingual contrastive learning baseline are marginal, and no statistically significant difference is observed between these two approaches.

The same pattern is observed when the basic contrastive learning baseline and the explicitly cross-lingual strategies are compared under the same hyperparameter settings. Neither approach consistently obtains lower PER across configurations: in some settings the basic baseline obtains lower PER, whereas in others the explicitly cross-lingual strategy does. Together with the significance-testing results, this provides no evidence that explicitly prioritizing cross-lingual positive pairs improves recognition over basic same-symbol contrastive learning.

Additionally, the best bilingual systems do not outperform the best monolingual contrastive learning systems when Dutch and English are evaluated separately. The bilingual setup looks useful when compared to a bilingual CTC-only baseline, but it does not show evidence that cross-lingual training has any clear advantage over training separate language-specific models.

Overall, the recognition results therefore support a benefit of phoneme-level contrastive learning over the bilingual CTC-only baseline in this case study, but do not provide evidence that explicitly modelling cross-lingual phoneme relationships yields an additional recognition improvement.

The embedding-space results show clearer effects of the bilingual contrastive objectives than the recognition results. Contrastive learning systems obtain lower cosine distances between same-phoneme representations and higher silhouette scores than CTC-only fine-tuning. The models trained with weighted loss achieve the highest silhouette scores, obtain lower cosine distances between cross-lingual equivalent phonemes and cross-lingual same-symbol phonemes, and show higher nearest-neighbor proportions for cross-lingual phoneme representations.

These results indicate that the bilingual contrastive objectives, particularly the weighted variants, shape the representation space in the intended cross-lingual direction. However, this structure remains incomplete. The equivalence-aware silhouette scores are generally lower than the general silhouette scores, indicating that grouping manually defined English-Dutch equivalents together does not produce stronger global separation according to the silhouette metric.

Overall, the representation space results show that explicitly modelling cross-lingual phoneme relationships can alter the learned geometry, but these changes are not accompanied by clear recognition improvements over basic phoneme-level contrastive learning. This suggests that stronger cross-lingual organization of the representation space is not sufficient on its own to improve PER.

6.2. SRQ1: HOW DO DIFFERENT TRIPLET SAMPLING STRATEGIES AFFECT RECOGNITION PERFORMANCE AND REPRESENTATION STRUCTURE?

To answer this question, three triplet sampling strategies were compared: basic same-symbol sampling, equivalence-prioritized bilingual sampling, and hierarchical bilingual sampling.

In terms of recognition performance, the basic same-symbol strategy provides a strong baseline. It obtains the lowest observed overall PER among the three unweighted sampling strategies, while the explicitly bilingual strategies produce very similar values and do not differ significantly from the basic strategy. These results therefore do not provide evidence that explicitly prioritizing cross-lingual positive relationships improves overall recognition beyond same-symbol contrastive learning.

The hierarchical strategy shows a somewhat more targeted numerical pattern. It obtains lower observed PER than the basic strategy across the equivalence-based evaluation scopes, including the overall equivalence subset. However, the tested difference for overall equivalence PER is not statistically significant. The results therefore suggest a possible concentration of the effect on the phoneme categories targeted by the hierarchical strategy, but do not provide sufficient evidence to conclude that hierarchical sampling improves recognition for these phonemes. Its overall PER also remains slightly higher than that of the basic strategy.

The representation space results similarly show little additional effect from changing the positive sampling strategy alone. Neither the equivalence-prioritized nor the hierarchical unweighted strategy obtains substantially lower cross-lingual cosine distances than basic same-symbol sampling, and their silhouette and nearest-neighbor results are also broadly similar. Thus, explicitly prioritizing cross-lingual positives during sampling

does not, by itself, produce a clear additional change in the representation space over basic contrastive learning.

One possible explanation is that the explicitly cross-lingual positive pairs make up only a limited part of the triplets contributing to the unweighted objective. Even when they are given priority during sampling, same-symbol triplets remain much more common, so the contribution of the explicitly bilingual relationships to the averaged loss may be limited. This interpretation is consistent with the stronger representation space changes observed when the cross-lingual triplet categories are given additional weight in the weighted-loss experiments.

6.3. SRQ2: HOW DOES WEIGHTING THE BILINGUAL CONTRASTIVE LOSS AFFECT RECOGNITION PERFORMANCE AND REPRESENTATION STRUCTURE COMPARED TO THE UNWEIGHTED OBJECTIVE?

As the previous section has shown, the sampling strategy itself does not have much of an effect on the performance or how the representation space is shaped. One possible reason for that is that the overall number of cross-lingual positives as opposed to same-language phonemes is too small to have an effect. Using weighted loss, it is possible to increase the influence of the triplets most directly related to the bilingual hypothesis. However, the effect of weighting differs between the equivalence-prioritized and hierarchical strategies.

For the equivalence-prioritized strategy, the weighted loss distinguishes only between the manually defined cross-lingual equivalents and the same-symbol phonemes, ignoring the differences between cross-lingual same-symbol phonemes and same-language same-symbol phonemes. In this scenario, the weighted configuration obtains numerically lower PER across several evaluation scopes, although the tested differences in overall and equivalence PER are not statistically significant. The embedding space shows clearer changes, through higher silhouette scores, higher nearest-neighbor proportions for cross-lingual equivalents, and lower cosine distances.

The loss used for the hierarchical strategy shows a slightly different pattern. In this case, the weighted loss separates three categories: manually defined cross-lingual different-symbol positives, cross-lingual same-symbol positives, and same-language same-symbol positives. Unlike in the equivalence-prioritized setup, the hierarchical weighted loss therefore explicitly distinguishes cross-lingual same-symbol positives from same-language same-symbol positives. This distinction is reflected particularly clearly in the embedding-space metrics: the weighted hierarchical approaches obtain the highest observed nearest-neighbor proportions for cross-lingual same-symbol phonemes, some of the highest silhouette scores, and lower cosine distances for the targeted cross-lingual relationships.

However, these representation space changes do not correspond to lower PER. The unweighted hierarchical model obtains lower overall PER than both weighted variants. The difference between the unweighted model and weighted Version 1 is not statistically significant, whereas weighted Version 2 produces a statistically significant increase in overall PER. The weighted hierarchical variants also obtain higher observed equivalence PER

than the unweighted model, although these differences are not statistically significant.

These results suggest that prioritizing cross-lingual relationships through weighting needs to be done carefully. While increasing their contribution can make the intended bilingual structure more visible in the representation space, giving them too much emphasis relative to same-language same-symbol positives may be detrimental to recognition. In particular, the results suggest that same-language same-symbol phonemes should not be overly downweighted in favour of cross-lingual positives.

The results therefore show that weighting can shape the representation space in the intended direction, but that the choice of individual weights is important. A representation space that more strongly reflects the manually designed bilingual hierarchy is not necessarily the representation space that produces the best phoneme recognition performance.

It is also worth noting that this experimental setup paired a specific sampling strategy with a corresponding weighting strategy, without evaluating all possible sampling-strategy and weighted-loss combinations. Although it is intuitive to pair a loss that distinguishes cross-lingual and same-language same-symbol positives with a sampling strategy that makes the same distinction, the two mechanisms may interact in ways that are not captured by the present experiments. Additional experiments would therefore be required to determine the specific contribution of the weighting strategy independently of the sampling procedure.

6.4. SRQ3: HOW DOES INCORPORATING TYPICAL SPEECH DATA AFFECT RECOGNITION PERFORMANCE AND REPRESENTATION STRUCTURE?

To answer this question, two setups were evaluated: a matched-size typical-dysarthric condition and a typical-data addition condition. The matched-size condition tested whether typical speech can replace part of the dysarthric training data, while the addition condition examined whether typical speech is beneficial when added on top of the full dysarthric training set.

The matched-size condition performs significantly worse than the dysarthric-data-only condition. This indicates that, in this personalized ASR setup, typical speech is not an effective substitute for target-speaker dysarthric speech. Replacing dysarthric samples with typical speech removes training examples that are directly matched to the target speaker and test condition. Although typical speech can provide additional examples of the relevant phonetic units, it does not contain the same speaker-specific dysarthric characteristics. The deterioration in recognition performance therefore emphasizes the importance of retaining target-speaker dysarthric data when the total amount of training data is fixed.

The typical-data addition condition shows the opposite effect, obtaining significantly lower PER than the dysarthric-data-only condition. However, this setup also increases the total amount of training data, so the improvement cannot be attributed specifically to the use of typical speech. It may instead result partly or entirely from the larger training set. Increasing the available training data from approximately 17 to 21 hours is associated with a relative reduction of approximately 14% in overall PER, highlighting the potential

value of additional training material in this low-data personalized setting.

The embedding-space results also differ from those of the dysarthric-data-only condition. In both typical-data setups, the cross-lingual pair categories obtain lower average cosine distances than same-phoneme same-language pairs, and cross-lingual categories generally appear more frequently among the nearest neighbors than in the dysarthric-data-only setting. Importantly, however, similar representation space patterns are observed in the two typical-data conditions despite their opposite effects on recognition performance.

Overall, the typical-data experiments show that the amount and composition of the training data have a substantial effect on recognition performance. Replacing target-speaker dysarthric speech with typical speech significantly worsens performance, whereas adding typical speech on top of the full dysarthric training set significantly improves it. At the same time, the two conditions produce broadly similar changes in cross-lingual representation structure. This further shows that stronger cross-lingual organization of the embedding space does not by itself determine recognition performance. The conclusions regarding the contrastive strategies therefore remain similar: explicitly bilingual objectives can shape the representation space, but they do not provide a clear recognition advantage over simpler phoneme-level contrastive learning.

6.5. THE IMPACT OF HYPERPARAMETERS

The contrastive learning experiments varied two main experimental dimensions across multiple conditions: the triplet loss weight λ_{triplet} and the alignment strategy used for phoneme-span extraction. The tested triplet loss weights were 0.05, 0.1, 0.2, 0.3, and 0.4, and the two alignment strategies were non-blank CTC spans and contiguous phoneme spans.

The results do not show a single triplet loss weight that consistently performs best across all conditions. Instead, the best value of λ_{triplet} differs depending on the training setup, sampling strategy, evaluation scope, and alignment type. This suggests that the contribution of the contrastive loss must be balanced carefully against the CTC objective. A weak contrastive weight may not sufficiently influence the representation space, while a stronger weight may interfere with the CTC objective or overemphasize noisy triplet relationships.

The alignment strategy also does not show a consistently dominant pattern. In some conditions, non-blank spans perform better, while in others, contiguous spans produce better results. This suggests that neither span definition is universally superior. Non-blank spans may provide cleaner phoneme-specific representations by excluding CTC blank regions, while contiguous spans may capture transitional acoustic information between phonemes. The fact that both strategies perform similarly in many cases suggests that the overall effect of alignment type is limited compared with the broader training condition.

The sensitivity to hyperparameters is another reason why the results should be interpreted cautiously. Since the best-performing configuration differs across conditions, the observed improvements may depend partly on the interaction between sampling strategy, alignment definition, and contrastive loss weight. This makes it difficult to attribute performance differences to a single methodological factor.

6.6. LIMITATIONS

Although the proposed contrastive learning approach affects the learned representation space, its impact on recognition performance remains limited. The aspects mentioned below help explain why:

Negative pair handling. The handling of positive pairs in this thesis is quite detailed, with different sampling strategies and weighted losses. In contrast, the negative samples are selected in a very simple way: through phonemes with labels different than the anchor and the positive (unless they are determined as a cross-lingual equivalent). The negatives are not graded by phonetic similarity, acoustic similarity, or model confusability. As a result, the contrastive loss will not focus on the phoneme contrasts that are most difficult for the recognizer.

Contrastive learning, as much as it aims to bring positives together, also works in a discriminative fashion, but it has not been explored. Recent contrastive learning approaches often use more structured or dynamic negative sampling strategies. However, the main focus of this thesis was exploring bilinguality within contrastive learning. Such bilingual relationships between the anchor and the negatives are extremely hard to define. While incorporating confusion-aware or dynamically selected hard negatives was possible, it was considered to be out of scope of this work and would therefore be a separate extension of the method.

Determining positive pairs. The cross-lingual positive pairs are based on manually defined phonetic equivalence between English and Dutch phonemes. This mapping is linguistically motivated, but it does not directly measure acoustic similarity in the actual recordings. This is especially important for dysarthric speech, since a dysarthric speaker might not pronounce phonemes in the same way as expected from canonical phonetic descriptions.

Therefore, some phoneme pairs treated as positives may not be acoustically close for this specific speaker. This may make the contrastive signal noisy: the loss encourages the model to bring certain phoneme pairs closer together, but these pairs may not always correspond to similar acoustic realizations in the data.

Dataset size and speaker coverage. The dysarthric experiments are based on a single dysarthric speaker. This is appropriate for a personalized ASR case study, but it limits the generalizability of the findings. The observed effects may depend on this speaker's articulation patterns, language background, severity of dysarthria, and recording conditions.

The small dataset size also limits how reliably the model can learn cross-lingual phoneme relationships. Some phoneme categories and equivalence pairs occur relatively rarely, which makes both training and evaluation more sensitive to individual examples.

Phoneme alignment extraction. The contrastive loss relies on phoneme-span embeddings extracted using precomputed CTC forced alignments. If these alignments contain errors, the extracted span embeddings may not correspond cleanly to the

intended phonemes. This can introduce noise into the triplet loss, since the model may be encouraged to compare embeddings that do not accurately represent the target phoneme categories.

This limitation is difficult to avoid in the absence of manual frame-level phoneme alignments. Dynamic alignments were also explored during method development, but they did not improve performance. Therefore, fixed CTC-derived alignments were used as a practical compromise, even though they may still introduce alignment noise.

Hyperparameter selection. The best-performing configurations reported in the main results and embedding-space analyses were selected according to the lowest observed test-set PER. This allowed the representation analysis to examine the embedding structure associated with the strongest observed recognition performance for each approach. However, this means the test set was used both for configuration selection and final performance evaluation, rather than being kept fully independent of model selection. Since multiple configurations were compared, selecting the minimum observed test PER may lead to an optimistic estimate of performance and makes very small differences between the best configurations particularly difficult to interpret. The complete results across all tested configurations are reported in the Appendix, but future experiments should select hyperparameters exclusively on the validation set and use the test set only for the final evaluation of the selected configuration.

Canonical phoneme labels. The phoneme labels used in this study were generated automatically from the orthographic transcriptions using *espeak-ng*. As a result, they represent canonical pronunciations rather than phonetic annotations of how the target speaker actually produced the words. Since phoneme substitutions are very common in dysarthric speech, this method of obtaining phoneme labels can introduce inaccuracies between actual phonemes pronounced and the generated labels. These labels are used to obtain phoneme alignments and determine the phoneme categories used in contrastive learning, and such mismatches between canonical labels and the speech signal may introduce additional noise into the training process.

6.7. REFLECTION: SCIENTIFIC, ETHICAL, AND SOCIAL IMPLICATIONS

This thesis was motivated by the limited performance of speaker-general ASR on dysarthric speech and by the scarcity of speaker-specific dysarthric training data. Speech produced by the same target speaker in another language offers an additional source of labeled data while preserving speaker-specific and dysarthric speech characteristics. The proposed phoneme-level contrastive learning method investigated whether relationships between Dutch and English phonemes could be used to exploit this additional speech more effectively. The mixed outcome of these experiments, with clearer cross-lingual organization of the representation space but no corresponding recognition advantage, makes the assumptions used to define these relationships particularly important to examine.

One aspect of the method that deserves particular reflection is the definition of cross-lingual positive pairs. The representation space changed more clearly than the recognition performance, which suggests that the relationships encouraged by the contrastive objective may not have aligned closely enough with the distinctions that were most relevant for the target speaker. In this study, these relationships were defined using canonical English–Dutch phonetic similarities. Using canonical relationships gave the study a consistent, linguistically informed basis for defining cross-lingual positive pairs. However, canonical similarity does not necessarily match how the target speaker produces the corresponding sounds. Dysarthria may alter the acoustic relationships between phonemes, and these speaker-specific patterns are not captured by the mapping used in this thesis. An alternative would be to derive positive relationships from the target speaker's own speech, for example using acoustic similarity between phoneme representations or the phoneme confusions made by the recognizer. This would address a different question: whether relationships derived from the speaker's actual productions are more useful than textbook linguistic relationships. The present study evaluates the latter approach, while the results motivate investigating the former in future work.

The use of a single target speaker also affects the interpretation of the results. Including several speakers would make it possible to test whether the findings generalize across speakers with dysarthria. However, it would also introduce differences in articulation, dysarthria severity, language background, and individual error patterns. These differences could make the effect of the proposed cross-lingual relationships harder to isolate. Focusing on one speaker is therefore consistent with the personalized setting of this thesis and allows the method to be studied without inter-speaker variation. The limitation is that the results describe one case and cannot be assumed to generalize to other speakers.

Two other methodological choices concern the construction of the contrastive examples. First, phoneme spans were obtained using automatic CTC-derived alignments rather than manually annotated boundaries. Manual annotation could provide more accurate phoneme spans, but would add substantial expert annotation requirements to a setting that is already constrained by limited speaker-specific dysarthric data. Second, negative sampling was kept simple rather than selecting acoustically similar or frequently confused phonemes as hard negatives. More targeted negatives could connect the contrastive objective more directly to recognition errors. However, changing the negative sampling strategy at the same time as the bilingual positive sampling strategy would introduce an additional experimental factor and make the effect of the cross-lingual positives harder to isolate.

These results are also important from an accessibility perspective. Speech-based technologies are increasingly used for communication and everyday tasks, yet systems trained primarily on typical speech may perform poorly for users with dysarthria because their speech can differ substantially from the speech represented during training. For some people with mobility limitations, voice-based interfaces may also provide an important alternative means of interacting with technology. The matched-size typical-speech experiment illustrates the importance of dysarthric training data: when the amount of data was kept approximately constant, replacing dysarthric speech with typical speech resulted in worse recognition. This suggests that readily available typical speech is not an equivalent substitute for dysarthric speech, and that improving dysarthric ASR requires methods

that make effective use of the comparatively scarce dysarthric data that is available.

7

Conclusion and Future Work

This thesis investigated whether bilingual phoneme-level contrastive learning can improve personalized dysarthric ASR by exploiting phonetic relationships between Dutch and English spoken by the same individual. In particular, it explored whether English and Dutch phoneme representations could reinforce one another during training by treating cross-lingual phoneme pairs as positives in a contrastive objective. Different positive sampling strategies and weighted contrastive losses were evaluated, together with two conditions incorporating additional typical speech data.

Across all results, a consistent pattern emerged: bilingual phoneme-level contrastive learning does measurably reshape the representation space in the intended direction, through reducing cosine distance between same-phoneme and cross-lingual equivalent embeddings and increasing silhouette scores relative to CTC-only fine-tuning, however, this restructuring does not translate into a consistent, statistically significant improvement in phoneme error rate over a simpler bilingual contrastive baseline. Basic same-symbol sampling already captured most of the achievable benefit; the more elaborate equivalence-prioritized and hierarchical strategies produced clearer or more targeted embedding structure without a corresponding general PER gain. In the dysarthric-data-only experiments, weighting could further strengthen the intended representation space effects, but overly downweighting same-language positives in the hierarchical objective hurt recognition performance. Typical speech data behaved differently depending on how it was used: substituting it for dysarthric speech hurt performance, while adding it on top of the full dysarthric set helped, though this gain is confounded with the resulting increase in total training data and cannot be cleanly attributed to typical speech itself. Results were also sensitive to hyperparameter selection throughout, with no single triplet-loss weight or alignment strategy performing best across all three data settings.

Overall, these results do not support the hypothesis that encoding canonical cross-lingual phoneme relationships is, by itself, sufficient to improve recognition in this personalized, single-speaker setting. This is not evidence that the bilingual contrastive objective had no effect, however: it demonstrably reshaped the representation space as intended, ruling out only the straightforward version of the hypothesis – that encoding textbook

cross-lingual phonetic equivalence as positives would improve recognition by itself. This narrows the search space for future work: the central challenge is not encouraging cross-lingual structure in the representation space, which the objective already achieves, but identifying cross-lingual relationships that reflect how the target speaker actually pronounces the corresponding phonemes. Additionally, the typical-data experiments, even though not the central focus of the thesis, provide a clear demonstration of how severely data scarcity constrains personalized dysarthric ASR.

The limitations of this work shed light on the reasons for lack of additional recognition improvements from the explicitly cross-lingual objectives. The cross-lingual equivalence mapping was derived from canonical English and Dutch phonetic relationships rather than the target speaker's actual acoustic realizations, and canonical similarity may not describe which sounds this individual produces similarly. Negative sampling was also kept comparatively simple, without regard to phonetic or acoustic similarity, so the contrastive objective was never given the opportunity to focus on the phoneme confusions that actually drive recognition errors – it is possible that the lack of PER improvement reflects this shallow negative sampling strategy rather than a limit of the bilingual hypothesis itself. Both limitations point to the same conclusion: the present results do not show that cross-lingual contrastive learning cannot help personalized dysarthric ASR, but rather that it needs to draw on more speaker-specific information when constructing triplets.

More broadly, the results highlight two observations about representation learning for ASR. First, a representation space that appears better organized according to embedding-space metrics is not necessarily more useful for recognition. In this thesis, stronger phoneme separation and closer targeted cross-lingual relationships were not consistently accompanied by lower PER. A similar disconnect was reported by Whetten et al. [112], who found that improved cluster separation did not necessarily correspond to better downstream ASR performance. Second, a more complex representation-learning objective does not necessarily produce better recognition performance. The explicitly bilingual and weighted objectives produced more targeted changes in the representation space, but these changes did not consistently improve PER over simpler contrastive learning. Równicka et al. [113] report a related result in ASR adaptation, where a multi-layer adaptation network was no more effective than a single linear adaptation layer. Together, these findings suggest that representation-learning methods should be evaluated primarily by whether the structure they impose is useful for the downstream recognition task, rather than by representation space metrics or methodological complexity alone.

7.1. FUTURE WORK

From the limitations discussed above, future work can explore the following directions:

Speaker-specific contrastive pairs. The cross-lingual positives in this study were defined using canonical English-Dutch phonetic equivalences, which may not reflect the actual acoustic realizations of a dysarthric speaker. Future approaches could instead select positives based on acoustic similarity or learned speaker-specific representations.

Negative sampling should be made more structured, prioritizing phonemes that are

acoustically close or frequently confused by the model, rather than any phoneme with a different label.

Typical speech data. Replacing dysarthric speech with typical speech degraded performance, whereas adding typical speech improved it; however, the latter also increased the total amount of training data. To determine to which extent typical speech data helps beyond simply providing more training examples, more experiments should be conducted with varying proportions of additional typical versus additional dysarthric speech.

Manually annotated phoneme labels. Future work could investigate whether using manually annotated phonetic labels instead of phoneme sequences generated from orthographic transcriptions leads to more reliable phoneme-level contrastive learning. Such annotations could more closely reflect the target speaker's actual speech production and would allow the effect of speaker-specific phoneme information to be separated from possible mismatches introduced by canonical G2P transcriptions.

Multi-speaker evaluation. The proposed approach should also be evaluated in a multi-speaker setting to determine whether the effects observed in this single-speaker case study generalize to a multi-speaker setting. Ideally, this would require a dataset containing multiple dysarthric speakers producing speech in more than one language, however, such multilingual dysarthric datasets are currently very limited. A more realistic alternative would be to use dysarthric speech datasets from different languages, even if the recordings come from different speakers. Although this would differ from the personalized bilingual setting considered in this thesis, it could help determine whether cross-lingual phoneme relationships provide useful information when inter-speaker variability is introduced.

DISCLAIMER ON THE USE OF AI

When working on this thesis, I used AI-based tools to support coding and writing. These tools were not used to replace my own research work, analysis, or conclusions. All AI outputs were carefully reviewed, and modified when necessary, before being included in the final output. No private, sensitive, or confidential data was uploaded to external AI tools.

CODING

ChatGPT was used for writing boilerplate code, debugging and improving code readability.

WRITING

ChatGPT and Claude were used to assist with the writing and editing process of this thesis, to improve grammar, clarity and sentence structure.

LATEX

ChatGPT was used for table formatting in LaTeX and adapting references to bibtex format.

BIBLIOGRAPHY

[1] Z. Qian and K. Xiao. "A Survey of Automatic Speech Recognition for Dysarthric Speech". In: *Electronics* 12.20 (2023). DOI: 10.3390/electronics12204278. URL: https://doi.org/10.3390/electronics12204278.

[2] S. Dickson et al. "Patients' experiences of disruptions associated with post-stroke dysarthria". In: *International Journal of Language & Communication Disorders* 43.2 (2008), pp. 135–153. DOI: 10.1080/13682820701862228. URL: https://doi.org/10.1080/13682820701862228.

[3] Neethu Mariam Joy and S. Umesh. "Improving Acoustic Models in TORGO Dysarthric Speech Database". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26.3 (2018), pp. 637–645. DOI: 10.1109/TNSRE.2018.2802914.

[4] Abner Hernandez et al. *Cross-lingual Self-Supervised Speech Representations for Improved Dysarthric Speech Recognition*. 2022. arXiv: 2204.01670 [cs.CL]. URL: https://arxiv.org/abs/2204.01670.

[5] Hillel Aryeh Steinmetz. "Transfer Learning Using L2 Speech to Improve Automatic Speech Recognition of Dysarthric Speech". Master's thesis. Seattle, Washington: University of Washington, Aug. 2023. URL: http://hdl.handle.net/1773/50477.

[6] Yuki Takashima, Tetsuya Takiguchi, and Yasuo Ariki. "End-to-end Dysarthric Speech Recognition Using Multiple Databases". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019, pp. 6395–6399. DOI: 10.1109/ICASSP.2019.8683803.

[7] Saierdaer Yusuyin et al. "Whistle: Data-Efficient Multilingual and Crosslingual Speech Recognition via Weakly Phonetic Supervision". In: *IEEE Transactions on Audio, Speech and Language Processing* 33 (2025), pp. 1440–1453. DOI: 10.1109/TASLPRO.2025.3550683.

[8] Piotr Żelasko et al. *That Sounds Familiar: an Analysis of Phonetic Representations Transfer Across Languages*. 2020. arXiv: 2005.08118 [eess.AS]. URL: https://arxiv.org/abs/2005.08118.

[9] Lidan Wu et al. "A Sequential Contrastive Learning Framework for Robust Dysarthric Speech Recognition". In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, pp. 7303–7307. DOI: 10.1109/ICASSP39728.2021.9415017.

[10] Shiyao Wang et al. "Enhancing Dysarthric Speech Recognition for Unseen Speakers via Prototype-Based Adaptation". In: *Interspeech 2024*. interspeech$_2$024. ISCA, 2024, pp. 1305–1309. DOI: 10.21437/interspeech.2024-1360. URL: http://dx.doi.org/10.21437/Interspeech.2024-1360.

[11] Wonjun Lee et al. *DyPCL: Dynamic Phoneme-level Contrastive Learning for Dysarthric Speech Recognition*. 2025. arXiv: 2501.19010 [cs.CL]. URL: https://arxiv.org/abs/2501.19010.

[12] Vivek Bhardwaj et al. "Automatic Speech Recognition (ASR) Systems for Children: A Systematic Literature Review". In: *Applied Sciences* 12 (Apr. 2022). DOI: 10.3390/app12094419.

[13] Bo Li et al. *Towards Fast and Accurate Streaming End-to-End ASR*. 2020. arXiv: 2004.11544 [eess.AS]. URL: https://arxiv.org/abs/2004.11544.

[14] Jinyu Li. *Recent Advances in End-to-End Automatic Speech Recognition*. 2022. arXiv: 2111.01690 [eess.AS]. URL: https://arxiv.org/abs/2111.01690.

[15] Alex Graves et al. "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks". In: vol. 2006. Jan. 2006, pp. 369–376. DOI: 10.1145/1143844.1143891.

[16] Abdelrahman Mohamed et al. "Self-Supervised Speech Representation Learning: A Review". In: *IEEE Journal of Selected Topics in Signal Processing* 16.6 (Oct. 2022), pp. 1179–1210. ISSN: 1941-0484. DOI: 10.1109/jstsp.2022.3207050. URL: http://dx.doi.org/10.1109/JSTSP.2022.3207050.

[17] Xiao Liu et al. "Self-Supervised Learning: Generative or Contrastive". In: *IEEE Transactions on Knowledge and Data Engineering* 35.1 (2023), pp. 857–876. DOI: 10.1109/TKDE.2021.3090866.

[18] Steffen Schneider et al. *wav2vec: Unsupervised Pre-training for Speech Recognition*. 2019. arXiv: 1904.05862 [cs.CL]. URL: https://arxiv.org/abs/1904.05862.

[19] Alexei Baevski et al. *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. 2020. arXiv: 2006.11477 [cs.CL]. URL: https://arxiv.org/abs/2006.11477.

[20] Arun Babu et al. *XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale*. 2021. arXiv: 2111.09296 [cs.CL]. URL: https://arxiv.org/abs/2111.09296.

[21] Panji Arisaputra and Amalia Zahra. "Indonesian Automatic Speech Recognition with XLSR-53". In: *Ingénierie des systèmes d information* 27.6 (Dec. 2022), pp. 973–982. ISSN: 2116-7125. DOI: 10.18280/isi.270614. URL: http://dx.doi.org/10.18280/isi.270614.

[22] Alexis Conneau et al. *Unsupervised Cross-lingual Representation Learning for Speech Recognition*. 2020. arXiv: 2006.13979 [cs.CL]. URL: https://arxiv.org/abs/2006.13979.

[23] Chih-Chen Chen et al. *Evaluating Self-Supervised Speech Representations for Indigenous American Languages*. 2023. arXiv: 2310.03639 [cs.CL]. URL: https://arxiv.org/abs/2310.03639.

[24] Dario Amodei et al. *Deep Speech 2: End-to-End Speech Recognition in English and Mandarin*. 2015. arXiv: 1512.02595 [cs.CL]. URL: https://arxiv.org/abs/1512.02595.

[25] W. Xiong et al. *Achieving Human Parity in Conversational Speech Recognition*. 2017. arXiv: 1610.05256 [cs.CL]. URL: https://arxiv.org/abs/1610.05256.

[26] Alec Radford et al. *Robust Speech Recognition via Large-Scale Weak Supervision*. 2022. arXiv: 2212.04356 [eess.AS]. URL: https://arxiv.org/abs/2212.04356.

[27] Yuanyuan Zhang et al. *Exploring data augmentation in bias mitigation against non-native-accented speech*. 2023. arXiv: 2312.15499 [eess.AS]. URL: https://arxiv.org/abs/2312.15499.

[28] Allison Koenecke et al. "Racial disparities in automated speech recognition". In: *Proceedings of the National Academy of Sciences* 117.14 (2020), pp. 7684–7689. DOI: 10.1073/pnas.1915768117. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1915768117. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1915768117.

[29] Joshua L Martin and Kelly Elizabeth Wright. "Bias in Automatic Speech Recognition: The Case of African American Language". In: *Applied Linguistics* 44.4 (Aug. 2023), pp. 613–630. ISSN: 0142-6001. DOI: 10.1093/applin/amac066. eprint: https://academic.oup.com/applij/article-pdf/44/4/613/51053335/amac066.pdf. URL: https://doi.org/10.1093/applin/amac066.

[30] Joshua Martin and Kevin Tang. "Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be"". In: Oct. 2020. DOI: 10.21437/Interspeech.2020-2893.

[31] Rachael Tatman and Conner Kasten. "Effects of Talker Dialect, Gender Race on Accuracy of Bing Speech and YouTube Automatic Captions". In: Aug. 2017, pp. 934–938. DOI: 10.21437/Interspeech.2017-1746.

[32] Hend ElGhazaly et al. *Exploring Gender Disparities in Automatic Speech Recognition Technology*. 2025. arXiv: 2502.18434 [cs.CL]. URL: https://arxiv.org/abs/2502.18434.

[33] Rachael Tatman. "Gender and Dialect Bias in YouTube's Automatic Captions". In: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Ed. by Dirk Hovy et al. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 53–59. DOI: 10.18653/v1/W17-1606. URL: https://aclanthology.org/W17-1606/.

[34] Lauren Werner, Gaojian Huang, and Brandon Pitts. "Automated Speech Recognition Systems and Older Adults: A Literature Review and Synthesis". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 63 (Nov. 2019), pp. 42–46. DOI: 10.1177/1071181319631121.

[35] May Pik Yu Chan et al. "Training and typological bias in ASR performance for world Englishes". In: Sept. 2022, pp. 1273–1277. DOI: 10.21437/Interspeech.2022-10869.

[36] Elsayed Issa, Mahmoud Ali, and Kevin Hirschi. “Measuring linguistic bias in ASR: Whisper large-v3 on non-native speech versus human perception”. In: *Procedia Computer Science* 275 (2026). 7th International Conference on AI in Computational Linguistics, pp. 692–699. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2026.01.080. URL: https://www.sciencedirect.com/science/article/pii/S1877050926000803.

[37] Tara N. Sainath et al. *A Streaming On-Device End-to-End Model Surpassing Server-Side Conventional Model Quality and Latency*. 2020. arXiv: 2003.12710 [cs.CL]. URL: https://arxiv.org/abs/2003.12710.

[38] Shahram Ghorbani and John H.L. Hansen. “Leveraging Native Language Information for Improved Accented Speech Recognition”. In: *Interspeech 2018*. interspeech$_2$018. ISCA, 2018, pp. 2449–2453. DOI: 10.21437/interspeech.2018-1378. URL: http://dx.doi.org/10.21437/Interspeech.2018-1378.

[39] Seyed Reza Shahamiri, Vanshika Lal, and Dhvani Shah. “Dysarthric Speech Transformer: A Sequence-to-Sequence Dysarthric Speech Recognition System”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2023), pp. 3407–3416. DOI: 10.1109/TNSRE.2023.3307020.

[40] Shujie HU et al. *On-the-fly Routing for Zero-shot MoE Speaker Adaptation of Speech Foundation Models for Dysarthric Speech Recognition*. 2025. arXiv: 2505.22072 [cs.SD]. URL: https://arxiv.org/abs/2505.22072.

[41] Márcio Fuckner et al. “Uncovering Bias in ASR Systems: Evaluating Wav2vec2 and Whisper for Dutch speakers”. In: *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)* (2023), pp. 146–151. URL: https://api.semanticscholar.org/CorpusID:265215536.

[42] Yuanyuan Zhang, Thomas De Valck, and Odette Scharenborg. “Speech recognition performance disparities between Dutch diverse speaker groups”. In: *Phonetica* (2026). DOI: doi:10.1515/phon-2025-0061.

[43] Siyuan Feng et al. “Towards inclusive automatic speech recognition”. In: *Computer Speech Language* 84 (Sept. 2023), p. 101567. DOI: 10.1016/j.csl.2023.101567.

[44] Sukairaj Hafiz Imam et al. *Automatic Speech Recognition (ASR) for African Low-Resource Languages: A Systematic Literature Review*. 2025. arXiv: 2510.01145 [cs.CL]. URL: https://arxiv.org/abs/2510.01145.

[45] Bao Thai et al. “Synthetic Data Augmentation for Improving Low-Resource ASR”. In: *2019 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*. 2019, pp. 1–9. DOI: 10.1109/WNYIPW.2019.8923082.

[46] Guanrou Yang et al. “Enhancing Low-Resource ASR through Versatile TTS: Bridging the Data Gap”. In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2025, pp. 1–5. DOI: 10.1109/ICASSP49660.2025.10889894.

7

[47] Claytone Sikasote and Antonios Anastasopoulos. “BembaSpeech: A Speech Recognition Corpus for the Bemba Language”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, June 2022, pp. 7277–7283. URL: https://aclanthology.org/2022.lrec-1.790/.

[48] Hemant Yadav and Sunayana Sitaram. “A Survey of Multilingual Models for Automatic Speech Recognition”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, June 2022, pp. 5071–5079. URL: https://aclanthology.org/2022.lrec-1.542/.

[49] Anjuli Kannan et al. *Large-Scale Multilingual Speech Recognition with a Streaming End-to-End Model*. 2019. arXiv: 1909.05330 [eess.AS]. URL: https://arxiv.org/abs/1909.05330.

[50] Shubham Toshniwal et al. *Multilingual Speech Recognition With A Single End-To-End Model*. 2018. arXiv: 1711.01694 [eess.AS]. URL: https://arxiv.org/abs/1711.01694.

[51] Pam Enderby. “Chapter 22 - Disorders of communication: dysarthria”. In: *Neurological Rehabilitation*. Ed. by Michael P. Barnes and David C. Good. Vol. 110. Handbook of Clinical Neurology. Elsevier, 2013, pp. 273–281. DOI: https://doi.org/10.1016/B978-0-444-52901-5.00022-8. URL: https://www.sciencedirect.com/science/article/pii/B9780444529015000228.

[52] Shansong Liu et al. “Recent Progress in the CUHK Dysarthric Speech Recognition System”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 2267–2281. ISSN: 2329-9304. DOI: 10.1109/taslp.2021.3091805. URL: http://dx.doi.org/10.1109/TASLP.2021.3091805.

[53] Kaila L. Stipancic et al. ““You Say Severe, I Say Mild”: Toward an Empirical Classification of Dysarthria Severity”. In: *Journal of Speech, Language, and Hearing Research* 64.12 (2021), pp. 4718–4735. DOI: 10.1044/2021_JSLHR-21-00197.

[54] Protima Nomo Sudro et al. “Significance of Data Augmentation for Improving Cleft Lip and Palate Speech Recognition”. In: *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 2021, pp. 484–490.

[55] Feifei Xiong, Jon Barker, and Heidi Christensen. “Phonetic Analysis of Dysarthric Speech Tempo and Applications to Robust Personalised Dysarthric Speech Recognition”. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019, pp. 5836–5840. DOI: 10.1109/ICASSP.2019.8683091.

[56] Bhavik Vachhani and Chitralekha Bhat. “Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition”. In: Sept. 2018, pp. 471–475. DOI: 10.21437/Interspeech.2018-1751.

[57] Daniel S. Park et al. “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition”. In: *Proceedings of Interspeech 2019*. ISCA, Sept. 2019, pp. 2613–2617. DOI: 10.21437/Interspeech.2019-2680. URL: https://doi.org/10.21437/Interspeech.2019-2680.

[58] Chitralekha Bhat and Helmer Strik. “Two-stage data augmentation for improved ASR performance for dysarthric speech”. In: *Computers in Biology and Medicine* 189 (2025), p. 109954. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2025.109954. URL: https://www.sciencedirect.com/science/article/pii/S0010482525003051.

[59] Mohammad Soleymanpour et al. “Synthesizing Dysarthric Speech Using Multi-Speaker TTS for Dysarthric Speech Recognition”. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. 2022, pp. 7382–7386. DOI: 10.1109/ICASSP43922.2022.9746585.

[60] Enno Hermann and Mathew Magimai-Doss. “Few-Shot Dysarthric Speech Recognition with Text-to-Speech Data Augmentation”. In: *Proceedings of Interspeech 2023*. Aug. 2023, pp. 156–160. DOI: 10.21437/Interspeech.2023-2481.

[61] John Harvill et al. “Synthesis of New Words for Improved Dysarthric Speech Recognition on an Expanded Vocabulary”. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, pp. 6428–6432. DOI: 10.1109/ICASSP39728.2021.9414869.

[62] Emre Yılmaz et al. “Articulatory Features for ASR of Pathological Speech”. In: Sept. 2018. DOI: 10.21437/Interspeech.2018-67.

[63] Bhavik Vachhani, Chitralekha Bhat, and Biswajit Das. “Deep Autoencoder Based Speech Features for Improved Dysarthric Speech Recognition”. In: Aug. 2017. DOI: 10.21437/Interspeech.2017-1318.

[64] Chitralekha Bhat et al. “Dysarthric Speech Recognition Using Time-Delay Neural Network Based Denoising Autoencoder”. In: *Proceedings of Interspeech 2018*. 2018, pp. 451–455. DOI: 10.21437/Interspeech.2018-1754. URL: https://doi.org/10.21437/Interspeech.2018-1754.

[65] Luke Prananta et al. *The Effectiveness of Time Stretching for Enhancing Dysarthric Speech for Improved Dysarthric Speech Recognition*. 2022. arXiv: 2201.04908 [cs.SD]. URL: https://arxiv.org/abs/2201.04908.

[66] Mengzhe Geng et al. *On-the-Fly Feature Based Rapid Speaker Adaptation for Dysarthric and Elderly Speech Recognition*. 2023. arXiv: 2203.14593 [eess.AS]. URL: https://arxiv.org/abs/2203.14593.

[67] Joel Shor et al. “Personalizing ASR for Dysarthric and Accented Speech with Limited Data”. In: *Interspeech 2019*. interspeech$_2$019. ISCA, 2019, pp. 784–788. DOI: 10.21437/interspeech.2019-1427. URL: http://dx.doi.org/10.21437/Interspeech.2019-1427.

[68] Jordan Green et al. “Automatic Speech Recognition of Disordered Speech: Personalized Models Outperforming Human Listeners on Short Phrases”. In: Aug. 2021, pp. 4778–4782. DOI: 10.21437/Interspeech.2021-1384.

[69] Yuanyuan Zhang et al. *Comparing Human and Automatic Recognition of Dutch Dysarthric Continuous Speech: A Case Study*. 2026. arXiv: 2606.30237 [cs.CL]. URL: https://arxiv.org/abs/2606.30237.

[70] Asahi Ogasawara et al. "SS-JDSC: Single-Speaker Japanese Dysarthric Speech Corpus". In: *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2026, pp. 19487–19491. DOI: 10.1109/ICASSP55912.2026.11464441.

[71] Emre Yılmaz et al. "Combining Non-Pathological Data of Different Language Varieties to Improve DNN-HMM Performance on Pathological Speech". In: *Proceedings of Interspeech 2016*. 2016, pp. 218–222. DOI: 10.21437/Interspeech.2016-109.

[72] Sofia Leivaditi et al. "Fine-Tuning Strategies for Dutch Dysarthric Speech Recognition: Evaluating the Impact of Healthy, Disease-Specific, and Speaker-Specific Data". In: *Proceedings of Interspeech 2024*. 2024, pp. 1295–1299. DOI: 10.21437/Interspeech.2024-1231.

[73] Yaroslav Ganin et al. *Domain-Adversarial Training of Neural Networks*. 2016. arXiv: 1505.07818 [stat.ML]. URL: https://arxiv.org/abs/1505.07818.

[74] Sining Sun et al. *Domain Adversarial Training for Accented Speech Recognition*. 2018. arXiv: 1806.02786 [cs.CL]. URL: https://arxiv.org/abs/1806.02786.

[75] Oliver Adams et al. "Massively Multilingual Adversarial Speech Recognition". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 96–108. DOI: 10.18653/v1/N19-1009. URL: https://aclanthology.org/N19-1009/.

[76] Dominika Woszczyk, Stavros Petridis, and David Millard. *Domain Adversarial Neural Networks for Dysarthric Speech Recognition*. 2020. arXiv: 2010.03623 [cs.SD]. URL: https://arxiv.org/abs/2010.03623.

[77] Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. "Contrastive Representation Learning: A Framework and Review". In: *IEEE Access* 8 (2020), pp. 193907–193934. DOI: 10.1109/ACCESS.2020.3031549.

[78] Ting Chen et al. *A Simple Framework for Contrastive Learning of Visual Representations*. 2020. arXiv: 2002.05709 [cs.LG]. URL: https://arxiv.org/abs/2002.05709.

[79] Prannay Khosla et al. *Supervised Contrastive Learning*. 2021. arXiv: 2004.11362 [cs.LG]. URL: https://arxiv.org/abs/2004.11362.

[80] Shipeng Liu et al. *Contrastive Learning for Image Complexity Representation*. 2024. arXiv: 2408.03230 [cs.CV]. URL: https://arxiv.org/abs/2408.03230.

[81] Olivier Henaff. “Data-Efficient Image Recognition with Contrastive Predictive Coding”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 4182–4192. URL: https://proceedings.mlr.press/v119/henaff20a.html.

[82] Kaiming He et al. *Momentum Contrast for Unsupervised Visual Representation Learning*. 2020. arXiv: 1911.05722 [cs.CV]. URL: https://arxiv.org/abs/1911.05722.

[83] Rui Zhang et al. “Contrastive Data and Learning for Natural Language Processing”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*. Ed. by Miguel Ballesteros, Yulia Tsvetkov, and Cecilia O. Alm. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 39–47. DOI: 10.18653/v1/2022.naacl-tutorials.6. URL: https://aclanthology.org/2022.naacl-tutorials.6/.

[84] Tianyu Gao, Xingcheng Yao, and Danqi Chen. *SimCSE: Simple Contrastive Learning of Sentence Embeddings*. 2022. arXiv: 2104.08821 [cs.CL]. URL: https://arxiv.org/abs/2104.08821.

[85] Zhuang Liu et al. *Contrastive Learning for Recommender System*. 2021. arXiv: 2101.01317 [cs.IR]. URL: https://arxiv.org/abs/2101.01317.

[86] Yanqiao Zhu et al. *An Empirical Study of Graph Contrastive Learning*. 2021. arXiv: 2109.01116 [cs.LG]. URL: https://arxiv.org/abs/2109.01116.

[87] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “FaceNet: A unified embedding for face recognition and clustering”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 815–823. DOI: 10.1109/cvpr.2015.7298682. URL: http://dx.doi.org/10.1109/CVPR.2015.7298682.

[88] Ling Song et al. “Visible-Infrared Cross-Modality Person Re-Identification via Adaptive Weighted Triplet Loss and Progressive Training”. In: *IEEE Access* 12 (2024), pp. 181799–181807. DOI: 10.1109/ACCESS.2024.3510425.

[89] Yu Xiong, Shixiong Xu, and Gaofeng Meng. “Distance-Ranking-Based Weighted Triplet Loss for Visual Place Recognition”. In: *2023 2nd International Conference on Artificial Intelligence, Human-Computer Interaction and Robotics (AIHCIR)*. 2023, pp. 94–100. DOI: 10.1109/AIHCIR61661.2023.00022.

[90] Jiantao Wu et al. *Rethinking Positive Pairs in Contrastive Learning*. 2025. arXiv: 2410.18200 [cs.CV]. URL: https://arxiv.org/abs/2410.18200.

[91] Zhikai Wang et al. “Relative Contrastive Learning for Sequential Recommendation with Similarity-based Positive Sample Selection”. In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. CIKM ’24. ACM, Oct. 2024, pp. 2493–2502. DOI: 10.1145/3627673.3679681. URL: http://dx.doi.org/10.1145/3627673.3679681.

[92] Vivien Cabannes et al. *Active Self-Supervised Learning: A Few Low-Cost Relationships Are All You Need*. 2023. arXiv: 2303.15256 [cs.LG]. URL: https://arxiv.org/abs/2303.15256.

[93] Vlad Sobal et al. *𝕏-Sample Contrastive Loss: Improving Contrastive Learning with Sample Similarity Graphs*. 2024. arXiv: 2407.18134 [cs.CV]. URL: https://arxiv.org/abs/2407.18134.

[94] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. *Representation Learning with Contrastive Predictive Coding*. 2019. arXiv: 1807.03748 [cs.LG]. URL: https://arxiv.org/abs/1807.03748.

[95] Chaitanya Talnikar et al. "Joint Masked CPC And CTC Training For ASR". In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, pp. 3045–3049. DOI: 10.1109/ICASSP39728.2021.9414227.

[96] Lixu Sun, Nurmemet Yolwas, and Lina Jiang. "A Method Improves Speech Recognition with Contrastive Learning in Low-Resource Languages". In: *Applied Sciences* 13.8 (2023). ISSN: 2076-3417. DOI: 10.3390/app13084836. URL: https://www.mdpi.com/2076-3417/13/8/4836.

[97] Alex Xiao, Christian Fuegen, and Abdelrahman Mohamed. *Contrastive Semi-supervised Learning for ASR*. 2021. arXiv: 2103.05149 [cs.CL]. URL: https://arxiv.org/abs/2103.05149.

[98] M.A. Yusnita et al. "Phoneme-based or isolated-word modeling speech recognition system? An overview". In: *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*. 2011, pp. 304–309. DOI: 10.1109/CSPA.2011.5759892.

[99] International Phonetic Association. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press, 1999. ISBN: 978-0521637510. URL: https://www.cambridge.org.

[100] Saierdaer Yusuyin et al. "Pronunciation-Lexicon Free Training for Phoneme-Based Crosslingual ASR via Joint Stochastic Approximation". In: *IEEE Transactions on Audio, Speech and Language Processing* 34 (2026), pp. 272–284. DOI: 10.1109/TASLPRO.2025.3646003.

[101] Sunakshi Mehra, Virender Ranga, and Ritu Agarwal. "PhonoBiEmbedNet: A Phoneme and Bigram Embedding Framework for Low-Resource Spoken Word Recognition". In: *Circuits, Systems, and Signal Processing* (Jan. 2026), pp. 1–35. DOI: 10.1007/s00034-025-03463-5.

[102] Ruizhe Huang et al. *Less Peaky and More Accurate CTC Forced Alignment by Label Priors*. 2024. arXiv: 2406.02560 [eess.AS]. URL: https://arxiv.org/abs/2406.02560.

[103] Beverley Collins and {Inger M.} Mees. *The Phonetics of English and Dutch*. English. 5th ed. Opstilling: 801.4 col Løbe nr.: 040732. Netherlands: Brill, 2003. ISBN: 9004103406.

[104] Nelleke Oostdijk. “The Spoken Dutch Corpus: Overview and first evaluation”. In: *Proceedings of LREC-2000, Athens* 2 (Jan. 2000).

[105] Vassil Panayotov et al. “Librispeech: An ASR corpus based on public domain audio books”. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

[106] eSpeak NG Developers. *eSpeak NG: Text-to-Speech Synthesizer*. https://github.com/espeak-ng/espeak-ng. GitHub repository. 2026.

[107] Mirco Ravanelli et al. *SpeechBrain: A General-Purpose Speech Toolkit*. 2021. arXiv: 2106.04624 [eess.AS]. URL: https://arxiv.org/abs/2106.04624.

[108] Peter J. Rousseeuw. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis”. In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

[109] Mohammad Idhom et al. “Evaluating Clustering Metrics for Time-Series Data: A Power Test Approach”. In: *2025 IEEE 11th Information Technology International Seminar (ITIS)*. 2025, pp. 368–373. DOI: 10.1109/ITIS67966.2025.11308723.

[110] M. Bisani and H. Ney. “Bootstrap estimates for confidence intervals in ASR performance evaluation”. In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. 2004, pp. I–409. DOI: 10.1109/ICASSP.2004.1326009.

[111] Zhe Liu and Fuchun Peng. *Statistical Testing on ASR Performance via Blockwise Bootstrap*. 2020. arXiv: 1912.09508 [stat.ML]. URL: https://arxiv.org/abs/1912.09508.

[112] Ryan Whetten et al. *Towards Early Prediction of Self-Supervised Speech Model Performance*. 2025. arXiv: 2501.05966 [cs.SD]. URL: https://arxiv.org/abs/2501.05966.

[113] Joanna Rownicka, Peter Bell, and Steve Renals. “Embeddings for DNN Speaker Adaptive Training”. In: *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2019, pp. 479–486. DOI: 10.1109/ASRU46091.2019.9004028.

A

APPENDIX

A.1. FULL RESULTS FOR DYSARTHRIC DATA ONLY EXPERIMENTS

This section provides the full phoneme error rate results complementing the results presented in Section 5.1.1.

Table A.1 presents full results for monolingual experiments. Table A.2 presents full results for the bilingual CTC-only fine-tuning experiments. Table A.3 presents full results for the contrastive learning with basic sampling strategy experiments. Table A.4 presents full results for the unweighted equivalence-prioritized strategy, while Table A.5 presents the corresponding weighted-loss results. Table A.6 presents full results for the unweighted hierarchical strategy, while Tables A.7 and A.8 present the two hierarchical weighted-loss variants.

A

Table A.1: Phoneme error rate results for the dysarthric-data-only monolingual contrastive learning experiments across alignment types and triplet loss weights. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. The best available result in each evaluation scope is underlined. Equiv. PER denotes PER computed over the language-specific phoneme equivalence set.

		English		Dutch	
Alignment	λ_{triplet}	PER	Equiv. PER	PER	Equiv. PER
CTC-only baseline		**8.55** S: 4.76 I: 1.56 D: 2.24	**14.27** S: 10.19 I: 1.49 D: 2.58	**11.44** S: 5.78 I: 2.17 D: 3.49	**<u>10.91</u>** S: 6.85 I: 1.42 D: 2.63
Non-blank	0.05	**8.81** S: 4.76 I: 1.67 D: 2.38	**14.13** S: 10.60 I: 1.22 D: 2.31	**11.74** S: 6.10 I: 2.17 D: 3.47	**12.11** S: 7.50 I: 1.38 D: 3.23
	0.10	**8.58** S: 4.59 I: 1.64 D: 2.34	**13.59** S: 10.05 I: 1.63 D: 1.90	**12.07** S: 6.17 I: 2.44 D: 3.45	**12.59** S: 8.06 I: 1.81 D: 2.72
	0.20	**8.76** S: 4.85 I: 1.62 D: 2.29	**13.86** S: 10.19 I: 1.49 D: 2.17	**11.67** S: 6.06 I: 2.48 D: 3.13	**11.85** S: 7.76 I: 1.90 D: 2.20
	0.30	**8.49** S: 4.80 I: 1.49 D: 2.20	**13.72** S: 10.46 I: 1.22 D: 2.04	**11.89** S: 6.20 I: 2.23 D: 3.47	**12.20** S: 7.59 I: 1.68 D: 2.93
	0.40	**<u>8.23</u>** S: 4.65 I: 1.46 D: 2.13	**<u>12.36</u>** S: 9.10 I: 1.22 D: 2.04	**12.01** S: 6.23 I: 2.41 D: 3.38	**12.46** S: 7.59 I: 2.07 D: 2.80
Contiguous	0.05	**8.72** S: 4.73 I: 1.57 D: 2.43	**13.32** S: 9.24 I: 1.36 D: 2.72	**11.94** S: 6.07 I: 2.35 D: 3.53	**12.37** S: 7.93 I: 1.68 D: 2.76
	0.10	**8.50** S: 4.64 I: 1.58 D: 2.29	**14.67** S: 9.92 I: 1.90 D: 2.85	**12.18** S: 6.36 I: 2.28 D: 3.54	**12.20** S: 7.89 I: 1.47 D: 2.84
	0.20	**8.38** S: 4.48 I: 1.61 D: 2.29	**<u>12.36</u>** S: 9.10 I: 0.68 D: 2.58	**11.90** S: 6.19 I: 2.42 D: 3.29	**11.85** S: 7.84 I: 1.59 D: 2.41
	0.30	**8.47** S: 4.80 I: 1.60 D: 2.07	**14.40** S: 10.46 I: 1.63 D: 2.31	**<u>11.29</u>** S: 5.83 I: 2.09 D: 3.37	**12.28** S: 7.63 I: 1.64 D: 3.02
	0.40	**–** S: – I: – D: –	**–** S: – I: – D: –	**11.94** S: 6.21 I: 2.26 D: 3.47	**12.07** S: 7.63 I: 1.64 D: 2.80

Table A.2: Phoneme error rate results for the dysarthric-data-only bilingual CTC-only fine-tuning. Substitution, insertion, and deletion rates are denoted by S, I, and D, respectively.

Evaluation scope	PER (%)	S (%)	I (%)	D (%)
Overall	10.46	5.74	1.94	2.78
Dutch	12.30	6.48	2.35	3.47
English	8.88	5.11	1.59	2.18
Equivalence phonemes	12.81	8.38	1.85	2.59
English equivalences	12.63	8.46	2.18	1.98
Dutch equivalences	12.93	8.32	1.64	2.97

A

Table A.3: Phoneme error rate (%) for the dysarthric-data-only basic CL across alignment types and triplet loss weights. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. The best available result in each evaluation scope is underlined. Equiv. denotes evaluation over the phoneme equivalence set.

Alignment	λ_{triplet}	Overall	Dutch	English	Equiv.	EN equiv.	NL equiv.
Non-blank	0.05	**10.36**	**12.19**	**8.81**	**12.29**	**12.01**	**12.46**
		S: 5.61	S: 6.22	S: 5.09	S: 8.06	S: 8.19	S: 7.97
		I: 1.74	I: 2.07	I: 1.45	I: 1.59	I: 1.84	I: 1.42
		D: 3.02	D: 3.90	D: 2.27	D: 2.64	D: 1.98	D: 3.06
	0.10	**10.45**	**12.01**	**9.11**	**12.10**	**12.76**	**11.68**
		S: 5.81	S: 6.42	S: 5.29	S: 8.19	S: 9.08	S: 7.63
		I: 1.81	I: 2.17	I: 1.50	I: 1.59	I: 1.77	I: 1.47
		D: 2.83	D: 3.42	D: 2.32	D: 2.32	D: 1.91	D: 2.59
	0.20	**10.34**	**12.03**	**8.90**	**12.39**	**12.35**	**12.41**
		S: 5.71	S: 6.41	S: 5.10	S: 8.38	S: 8.87	S: 8.06
		I: 1.78	I: 2.17	I: 1.44	I: 1.51	I: 1.57	I: 1.47
		D: 2.86	D: 3.44	D: 2.36	D: 2.51	D: 1.91	D: 2.89
	0.30	**10.01**	**11.60**	**8.65**	**11.99**	**12.35**	**11.77**
		S: 5.61	S: 6.35	S: 4.98	S: 8.19	S: 8.60	S: 7.93
		I: 1.69	I: 1.99	I: 1.43	I: 1.40	I: 1.64	I: 1.25
		D: 2.70	D: 3.25	D: 2.24	D: 2.40	D: 2.12	D: 2.59
	0.40	**10.36**	**11.87**	**9.06**	**12.92**	**12.76**	**13.02**
		S: 5.83	S: 6.47	S: 5.28	S: 8.80	S: 9.15	S: 8.58
		I: 1.63	I: 1.96	I: 1.35	I: 1.48	I: 1.64	I: 1.38
		D: 2.90	D: 3.44	D: 2.43	D: 2.64	D: 1.98	D: 3.06
Contiguous	0.05	**10.68**	**12.19**	**9.39**	**11.94**	**12.42**	**11.64**
		S: 5.77	S: 6.26	S: 5.35	S: 8.06	S: 8.87	S: 7.54
		I: 1.76	I: 2.08	I: 1.49	I: 1.27	I: 1.43	I: 1.16
		D: 3.15	D: 3.85	D: 2.55	D: 2.62	D: 2.12	D: 2.93
	0.10	**10.42**	**12.27**	**8.83**	**12.44**	**11.95**	**12.76**
		S: 5.71	S: 6.40	S: 5.12	S: 8.22	S: 8.26	S: 8.19
		I: 1.71	I: 2.13	I: 1.36	I: 1.69	I: 1.71	I: 1.68
		D: 3.00	D: 3.74	D: 2.36	D: 2.54	D: 1.98	D: 2.89
	0.20	**9.96**	**11.46**	**8.68**	**11.78**	**11.67**	**11.85**
		S: 5.60	S: 6.27	S: 5.03	S: 8.08	S: 7.92	S: 8.19
		I: 1.65	I: 1.88	I: 1.45	I: 1.43	I: 1.91	I: 1.12
		D: 2.71	D: 3.32	D: 2.20	D: 2.27	D: 1.84	D: 2.54
	0.30	**10.16**	**11.79**	**8.77**	**12.47**	**12.70**	**12.33**
		S: 5.59	S: 6.18	S: 5.09	S: 8.14	S: 8.81	S: 7.72
		I: 1.68	I: 2.09	I: 1.33	I: 1.69	I: 1.84	I: 1.59
		D: 2.89	D: 3.52	D: 2.35	D: 2.64	D: 2.05	D: 3.02
	0.40	**10.23**	**11.69**	**8.99**	**12.07**	**12.70**	**11.68**
		S: 5.61	S: 6.17	S: 5.12	S: 8.11	S: 8.53	S: 7.84
		I: 1.68	I: 2.02	I: 1.39	I: 1.56	I: 1.98	I: 1.29
		D: 2.95	D: 3.49	D: 2.48	D: 2.40	D: 2.18	D: 2.54

Table A.4: Phoneme error rate (%) for the dysarthric-data-only unweighted equivalence-prioritized setup across alignment types and triplet loss weights. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. The best available PER in each evaluation scope is underlined. Equiv. denotes evaluation over the phoneme equivalence set.

Alignment	λ_{triplet}	Overall	Dutch	English	Equiv.	EN equiv.	NL equiv.
Non-blank	0.05	**10.33**	**11.88**	**9.00**	**12.29**	**12.01**	**12.46**
		S: 5.73	S: 6.22	S: 5.31	S: 8.45	S: 8.67	S: 8.32
		I: 1.87	I: 2.24	I: 1.55	I: 1.56	I: 1.71	I: 1.47
		D: 2.72	D: 3.41	D: 2.13	D: 2.27	D: 1.64	D: 2.67
	0.10	**10.15**	**11.69**	**8.84**	**12.13**	**12.63**	**11.81**
		S: 5.64	S: 6.19	S: 5.18	S: 8.19	S: 8.74	S: 7.84
		I: 1.78	I: 2.11	I: 1.50	I: 1.61	I: 1.91	I: 1.42
		D: 2.72	D: 3.38	D: 2.16	D: 2.32	D: 1.98	D: 2.54
	0.20	**10.43**	**12.10**	**9.01**	**12.50**	**12.70**	**12.37**
		S: 5.82	S: 6.40	S: 5.32	S: 8.32	S: 9.08	S: 7.84
		I: 1.97	I: 2.38	I: 1.61	I: 1.77	I: 1.91	I: 1.68
		D: 2.65	D: 3.32	D: 2.08	D: 2.40	D: 1.71	D: 2.84
	0.30	**10.46**	**11.88**	**9.25**	**12.05**	**12.70**	**11.64**
		S: 5.77	S: 6.18	S: 5.42	S: 8.08	S: 8.67	S: 7.72
		I: 1.86	I: 2.19	I: 1.58	I: 1.59	I: 1.98	I: 1.34
		D: 2.84	D: 3.51	D: 2.26	D: 2.38	D: 2.05	D: 2.59
Contiguous	0.05	**10.37**	**12.07**	**8.92**	**12.05**	**12.35**	**11.85**
		S: 5.71	S: 6.49	S: 5.04	S: 8.35	S: 8.53	S: 8.23
		I: 1.93	I: 2.34	I: 1.58	I: 1.48	I: 1.77	I: 1.29
		D: 2.73	D: 3.24	D: 2.29	D: 2.22	D: 2.05	D: 2.33
	0.10	**10.32**	**11.71**	**9.13**	**11.76**	**12.63**	**11.38**
		S: 5.74	S: 6.20	S: 5.35	S: 8.11	S: 8.81	S: 7.67
		I: 1.93	I: 2.28	I: 1.63	I: 1.51	I: 1.98	I: 1.21
		D: 2.65	D: 3.23	D: 2.16	D: 2.25	D: 1.84	D: 2.50
	0.20	**10.37**	**11.78**	**9.18**	**12.34**	**12.97**	**11.94**
		S: 5.70	S: 6.12	S: 5.33	S: 8.08	S: 8.81	S: 7.63
		I: 1.74	I: 2.02	I: 1.50	I: 1.64	I: 2.12	I: 1.34
		D: 2.94	D: 3.63	D: 2.34	D: 2.62	D: 2.05	D: 2.97
	0.30	**10.26**	**12.02**	**8.76**	**12.58**	**12.56**	**12.59**
		S: 5.70	S: 6.43	S: 5.07	S: 8.48	S: 8.67	S: 8.36
		I: 1.70	I: 2.10	I: 1.36	I: 1.59	I: 1.84	I: 1.42
		D: 2.86	D: 3.49	D: 2.32	D: 2.51	D: 2.05	D: 2.80
	0.40	**10.25**	**11.86**	**8.87**	**11.99**	**12.22**	**11.85**
		S: 5.64	S: 6.16	S: 5.20	S: 8.16	S: 8.81	S: 7.76
		I: 1.83	I: 2.28	I: 1.44	I: 1.43	I: 1.64	I: 1.29
		D: 2.78	D: 3.42	D: 2.23	D: 2.40	D: 1.77	D: 2.80

A

Table A.5: Phoneme error rate (%) for the dysarthric-data-only weighted equivalence-prioritized setup across alignment types and triplet loss weights. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. The best PER in each evaluation scope is underlined. Equiv. denotes evaluation over the phoneme equivalence set.

Alignment	λ_{triplet}	Overall	Dutch	English	Equiv.	EN equiv.	NL equiv.
Non-blank	0.05	**10.15**	**11.59**	**8.93**	**12.29**	**12.56**	**12.11**
		S: 5.59	S: 6.07	S: 5.17	S: 8.40	S: 8.87	S: 8.10
		I: 1.73	I: 2.05	I: 1.45	I: 1.53	I: 1.84	I: 1.34
		D: 2.84	D: 3.47	D: 2.31	D: 2.35	D: 1.84	D: 2.67
	0.10	**10.28**	**11.69**	**9.08**	**12.60**	**12.76**	**12.50**
		S: 5.68	S: 6.09	S: 5.33	S: 8.59	S: 8.81	S: 8.45
		I: 1.83	I: 2.20	I: 1.51	I: 1.51	I: 1.77	I: 1.34
		D: 2.77	D: 3.40	D: 2.24	D: 2.51	D: 2.18	D: 2.72
	0.20	**10.14**	**11.84**	**8.69**	**12.66**	**12.49**	**12.76**
		S: 5.61	S: 6.28	S: 5.03	S: 8.24	S: 8.60	S: 8.02
		I: 1.78	I: 2.14	I: 1.47	I: 1.69	I: 1.91	I: 1.55
		D: 2.75	D: 3.41	D: 2.19	D: 2.72	D: 1.98	D: 3.19
	0.30	**10.18**	**11.79**	**8.79**	**12.13**	**12.49**	**11.90**
		S: 5.58	S: 6.22	S: 5.03	S: 8.22	S: 8.94	S: 7.76
		I: 1.84	I: 2.15	I: 1.56	I: 1.64	I: 1.98	I: 1.42
		D: 2.76	D: 3.41	D: 2.20	D: 2.27	D: 1.57	D: 2.72
	0.40	**9.98**	**11.51**	**8.68**	**11.89**	**12.49**	**11.51**
		S: 5.56	S: 6.02	S: 5.17	S: 7.98	S: 8.60	S: 7.59
		I: 1.74	I: 2.16	I: 1.39	I: 1.51	I: 1.77	I: 1.34
		D: 2.68	D: 3.32	D: 2.12	D: 2.40	D: 2.12	D: 2.59
Contiguous	0.05	**10.16**	**11.84**	**8.73**	**11.94**	**11.67**	**12.11**
		S: 5.60	S: 6.34	S: 4.98	S: 8.22	S: 7.99	S: 8.36
		I: 1.70	I: 1.99	I: 1.45	I: 1.32	I: 1.57	I: 1.16
		D: 2.86	D: 3.50	D: 2.31	D: 2.40	D: 2.12	D: 2.59
	0.10	**10.21**	**11.66**	**8.97**	**12.07**	**12.42**	**11.85**
		S: 5.42	S: 5.87	S: 5.04	S: 7.98	S: 8.74	S: 7.50
		I: 1.73	I: 2.10	I: 1.42	I: 1.51	I: 1.71	I: 1.38
		D: 3.05	D: 3.69	D: 2.51	D: 2.59	D: 1.98	D: 2.97
	0.20	**9.96**	**11.64**	**8.52**	**11.97**	**12.08**	**11.90**
		S: 5.50	S: 6.19	S: 4.92	S: 8.40	S: 8.67	S: 8.23
		I: 1.70	I: 2.05	I: 1.39	I: 1.32	I: 1.71	I: 1.08
		D: 2.76	D: 3.40	D: 2.21	D: 2.25	D: 1.71	D: 2.59
	0.30	**10.49**	**11.92**	**9.27**	**12.50**	**13.11**	**12.11**
		S: 5.77	S: 6.24	S: 5.38	S: 8.35	S: 9.35	S: 7.72
		I: 1.79	I: 2.11	I: 1.51	I: 1.56	I: 1.71	I: 1.47
		D: 2.93	D: 3.57	D: 2.38	D: 2.59	D: 2.05	D: 2.93
	0.40	**9.94**	**11.39**	**8.70**	**12.07**	**12.42**	**11.85**
		S: 5.48	S: 5.98	S: 5.05	S: 8.27	S: 8.46	S: 8.15
		I: 1.79	I: 2.13	I: 1.51	I: 1.40	I: 1.91	I: 1.08
		D: 2.66	D: 3.27	D: 2.14	D: 2.40	D: 2.05	D: 2.63

Table A.6: Phoneme error rate (%) for the dysarthric-data-only unweighted hierarchical setup across alignment types and triplet loss weights. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. The best PER in each evaluation scope is underlined. Equiv. denotes evaluation over the phoneme equivalence set.

Alignment	λ_{triplet}	Overall	Dutch	English	Equiv.	EN equiv.	NL equiv.
Non-blank	0.05	**10.44**	**12.05**	**9.06**	**11.89**	**11.81**	**11.94**
		S: 5.83	S: 6.49	S: 5.27	S: 8.14	S: 8.60	S: 7.84
		I: 1.83	I: 2.29	I: 1.44	I: 1.29	I: 1.37	I: 1.25
		D: 2.78	D: 3.27	D: 2.36	D: 2.46	D: 1.84	D: 2.84
	0.10	**10.29**	**11.95**	**8.87**	**11.76**	**12.15**	**11.51**
		S: 5.76	S: 6.39	S: 5.23	S: 8.22	S: 8.87	S: 7.80
		I: 1.85	I: 2.24	I: 1.51	I: 1.37	I: 1.43	I: 1.34
		D: 2.68	D: 3.32	D: 2.13	D: 2.17	D: 1.84	D: 2.37
	0.20	**10.26**	**11.98**	**8.79**	**12.10**	**11.74**	**12.33**
		S: 5.73	S: 6.32	S: 5.22	S: 8.06	S: 7.78	S: 8.23
		I: 1.84	I: 2.31	I: 1.44	I: 1.61	I: 1.84	I: 1.47
		D: 2.69	D: 3.35	D: 2.13	D: 2.43	D: 2.12	D: 2.63
	0.30	**10.11**	**11.66**	**8.79**	**11.44**	**11.67**	**11.29**
		S: 5.60	S: 6.20	S: 5.08	S: 7.79	S: 8.05	S: 7.63
		I: 1.78	I: 2.02	I: 1.58	I: 1.37	I: 1.91	I: 1.03
		D: 2.73	D: 3.43	D: 2.13	D: 2.27	D: 1.71	D: 2.63
	0.40	**10.02**	**11.63**	**8.64**	**11.73**	**11.81**	**11.68**
		S: 5.48	S: 6.08	S: 4.97	S: 8.14	S: 8.33	S: 8.02
		I: 1.73	I: 2.08	I: 1.43	I: 1.29	I: 1.50	I: 1.16
		D: 2.80	D: 3.47	D: 2.24	D: 2.30	D: 1.98	D: 2.50
Contiguous	0.05	**10.31**	**11.47**	**9.31**	**11.23**	**11.54**	**11.21**
		S: 5.71	S: 6.01	S: 5.44	S: 7.77	S: 7.92	S: 7.67
		I: 1.69	I: 1.92	I: 1.49	I: 1.16	I: 1.57	I: 0.91
		D: 2.91	D: 3.53	D: 2.38	D: 2.40	D: 2.05	D: 2.63
	0.10	**10.25**	**11.65**	**9.06**	**11.99**	**12.42**	**11.72**
		S: 5.63	S: 6.16	S: 5.17	S: 8.22	S: 9.01	S: 7.72
		I: 1.74	I: 2.10	I: 1.43	I: 1.35	I: 1.50	I: 1.25
		D: 2.89	D: 3.39	D: 2.46	D: 2.43	D: 1.91	D: 2.76
	0.20	**10.42**	**12.00**	**9.08**	**11.99**	**12.22**	**11.85**
		S: 5.73	S: 6.31	S: 5.24	S: 7.98	S: 8.12	S: 7.89
		I: 1.88	I: 2.32	I: 1.51	I: 1.66	I: 1.91	I: 1.51
		D: 2.81	D: 3.38	D: 2.32	D: 2.35	D: 2.18	D: 2.46
	0.30	**10.21**	**11.73**	**8.91**	**11.99**	**12.08**	**11.94**
		S: 5.61	S: 6.04	S: 5.24	S: 8.08	S: 8.53	S: 7.80
		I: 1.74	I: 2.18	I: 1.35	I: 1.51	I: 1.77	I: 1.34
		D: 2.86	D: 3.50	D: 2.32	D: 2.40	D: 1.77	D: 2.80
	0.40	**10.13**	**11.65**	**8.83**	**12.21**	**12.35**	**12.11**
		S: 5.56	S: 5.93	S: 5.24	S: 8.38	S: 8.94	S: 8.02
		I: 1.81	I: 2.26	I: 1.42	I: 1.48	I: 1.71	I: 1.34
		D: 2.76	D: 3.45	D: 2.17	D: 2.35	D: 1.71	D: 2.76

A

Table A.7: Phoneme error rate (%) for the dysarthric-data-only weighted hierarchical setup ver. 1 across alignment types and triplet loss weights. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. The best available PER in each evaluation scope is underlined. Equiv. denotes evaluation over the phoneme equivalence set.

Alignment	λ_{triplet}	Overall	Dutch	English	Equiv.	EN equiv.	NL equiv.
Non-blank	0.05	**10.53**	**11.90**	**9.35**	**12.29**	**12.83**	**11.94**
		S: 5.67	S: 6.04	S: 5.36	S: 8.11	S: 8.67	S: 7.76
		I: 1.87	I: 2.20	I: 1.58	I: 1.48	I: 1.91	I: 1.21
		D: 2.99	D: 3.66	D: 2.41	D: 2.69	D: 2.25	D: 2.97
	0.10	**10.33**	**11.78**	**9.09**	**11.94**	**12.42**	**11.64**
		S: 5.73	S: 6.17	S: 5.36	S: 8.16	S: 9.08	S: 7.59
		I: 1.87	I: 2.27	I: 1.53	I: 1.64	I: 1.84	I: 1.51
		D: 2.73	D: 3.34	D: 2.20	D: 2.14	D: 1.50	D: 2.54
	0.20	**10.54**	**12.33**	**9.02**	**13.00**	**13.04**	**12.97**
		S: 5.90	S: 6.69	S: 5.22	S: 9.01	S: 9.08	S: 8.97
		I: 1.87	I: 2.27	I: 1.53	I: 1.69	I: 2.12	I: 1.42
		D: 2.77	D: 3.37	D: 2.26	D: 2.30	D: 1.84	D: 2.59
	0.30	**10.24**	**11.68**	**9.02**	**12.15**	**13.04**	**11.59**
		S: 5.56	S: 6.04	S: 5.15	S: 8.35	S: 9.15	S: 7.84
		I: 1.85	I: 2.14	I: 1.60	I: 1.43	I: 1.84	I: 1.16
		D: 2.83	D: 3.49	D: 2.27	D: 2.38	D: 2.05	D: 2.59
	0.40	**10.56**	**12.29**	**9.09**	**11.99**	**11.26**	**12.46**
		S: 5.87	S: 6.58	S: 5.26	S: 8.19	S: 7.85	S: 8.41
		I: 1.92	I: 2.31	I: 1.58	I: 1.56	I: 1.91	I: 1.34
		D: 2.78	D: 3.41	D: 2.24	D: 2.25	D: 1.50	D: 2.72
Contiguous	0.05	**10.90**	**12.34**	**9.67**	**12.07**	**12.76**	**11.64**
		S: 5.96	S: 6.42	S: 5.58	S: 8.35	S: 9.49	S: 7.63
		I: 1.96	I: 2.26	I: 1.69	I: 1.59	I: 1.84	I: 1.42
		D: 2.98	D: 3.66	D: 2.40	D: 2.14	D: 1.43	D: 2.59
	0.10	**10.89**	**12.52**	**9.50**	**12.73**	**13.04**	**12.54**
		S: 5.97	S: 6.52	S: 5.50	S: 8.75	S: 9.22	S: 8.45
		I: 1.88	I: 2.24	I: 1.57	I: 1.51	I: 1.91	I: 1.25
		D: 3.05	D: 3.77	D: 2.43	D: 2.48	D: 1.91	D: 2.84
	0.20	**10.57**	**12.25**	**9.14**	**12.39**	**11.74**	**12.80**
		S: 5.67	S: 6.27	S: 5.17	S: 8.43	S: 8.19	S: 8.58
		I: 1.87	I: 2.35	I: 1.47	I: 1.40	I: 1.37	I: 1.42
		D: 3.03	D: 3.64	D: 2.50	D: 2.56	D: 2.18	D: 2.80
	0.30	**10.46**	**12.01**	**9.14**	**11.97**	**12.63**	**11.55**
		S: 5.79	S: 6.30	S: 5.36	S: 8.32	S: 9.15	S: 7.80
		I: 1.89	I: 2.35	I: 1.50	I: 1.40	I: 1.64	I: 1.25
		D: 2.78	D: 3.36	D: 2.28	D: 2.25	D: 1.84	D: 2.50
	0.40	**10.65**	**12.31**	**9.24**	**12.55**	**12.49**	**12.59**
		S: 5.81	S: 6.34	S: 5.35	S: 8.59	S: 9.08	S: 8.28
		I: 1.83	I: 2.22	I: 1.49	I: 1.53	I: 1.57	I: 1.51
		D: 3.02	D: 3.74	D: 2.41	D: 2.43	D: 1.84	D: 2.80

Table A.8: Phoneme error rate (%) for the dysarthric-data-only weighted hierarchical setup ver. 2 across alignment types and triplet loss weights. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. The best available PER in each evaluation scope is underlined. Equiv. denotes evaluation over the phoneme equivalence set.

Alignment	λ_{triplet}	Overall	Dutch	English	Equiv.	EN equiv.	NL equiv.
Non-blank	0.05	**10.38**	**12.03**	**8.97**	**12.26**	**11.67**	**12.63**
		S: 5.65	S: 6.31	S: 5.09	S: 8.14	S: 7.85	S: 8.32
		I: 1.83	I: 2.21	I: 1.49	I: 1.43	I: 1.50	I: 1.38
		D: 2.90	D: 3.50	D: 2.39	D: 2.69	D: 2.32	D: 2.93
	0.10	**10.52**	**12.03**	**9.22**	**12.29**	**12.63**	**12.07**
		S: 5.72	S: 6.22	S: 5.30	S: 8.53	S: 9.42	S: 7.97
		I: 1.81	I: 2.14	I: 1.54	I: 1.45	I: 1.43	I: 1.47
		D: 2.98	D: 3.68	D: 2.38	D: 2.30	D: 1.77	D: 2.63
	0.20	**10.48**	**12.06**	**9.13**	**12.02**	**12.15**	**11.94**
		S: 5.78	S: 6.28	S: 5.35	S: 8.30	S: 8.67	S: 8.06
		I: 1.94	I: 2.39	I: 1.55	I: 1.53	I: 1.50	I: 1.55
		D: 2.77	D: 3.38	D: 2.24	D: 2.19	D: 1.98	D: 2.33
	0.30	**10.46**	**12.10**	**9.06**	**11.99**	**11.74**	**12.16**
		S: 5.74	S: 6.31	S: 5.25	S: 7.93	S: 8.12	S: 7.80
		I: 1.89	I: 2.29	I: 1.55	I: 1.69	I: 1.71	I: 1.68
		D: 2.84	D: 3.51	D: 2.27	D: 2.38	D: 1.91	D: 2.67
	0.40	**10.27**	**11.60**	**9.13**	**11.73**	**12.42**	**11.29**
		S: 5.65	S: 5.93	S: 5.42	S: 8.30	S: 9.42	S: 7.59
		I: 1.86	I: 2.24	I: 1.54	I: 1.51	I: 1.71	I: 1.38
		D: 2.75	D: 3.43	D: 2.17	D: 1.93	D: 1.30	D: 2.33
Contiguous	0.05	**10.63**	**11.92**	**9.54**	**12.89**	**13.17**	**12.72**
		S: 5.96	S: 6.27	S: 5.69	S: 9.14	S: 9.97	S: 8.62
		I: 1.82	I: 2.15	I: 1.54	I: 1.24	I: 1.43	I: 1.12
		D: 2.85	D: 3.50	D: 2.31	D: 2.51	D: 1.77	D: 2.97
	0.10	**10.59**	**12.36**	**9.08**	**12.05**	**11.81**	**12.20**
		S: 5.83	S: 6.44	S: 5.31	S: 8.11	S: 8.67	S: 7.76
		I: 1.88	I: 2.24	I: 1.58	I: 1.40	I: 1.37	I: 1.42
		D: 2.88	D: 3.68	D: 2.19	D: 2.54	D: 1.77	D: 3.02
	0.20	**10.85**	**12.41**	**9.52**	**12.73**	**12.83**	**12.67**
		S: 6.02	S: 6.47	S: 5.63	S: 8.67	S: 9.15	S: 8.36
		I: 1.86	I: 2.29	I: 1.49	I: 1.59	I: 1.77	I: 1.47
		D: 2.98	D: 3.65	D: 2.40	D: 2.48	D: 1.91	D: 2.84
	0.30	**10.68**	**12.38**	**9.24**	**12.13**	**11.47**	**12.54**
		S: 5.88	S: 6.57	S: 5.29	S: 8.32	S: 7.99	S: 8.53
		I: 1.90	I: 2.24	I: 1.61	I: 1.45	I: 1.77	I: 1.25
		D: 2.90	D: 3.57	D: 2.33	D: 2.35	D: 1.71	D: 2.76
	0.40	**10.45**	**11.96**	**9.16**	**12.42**	**12.76**	**12.20**
		S: 5.77	S: 6.37	S: 5.26	S: 8.32	S: 9.01	S: 7.89
		I: 1.83	I: 2.14	I: 1.56	I: 1.74	I: 1.84	I: 1.68
		D: 2.85	D: 3.45	D: 2.33	D: 2.35	D: 1.91	D: 2.63

A

A.2. FULL RESULTS FOR EXPERIMENTS INCLUDING ADDITIONAL TYPICAL DATA

A.2.1. MATCHED-SIZE TYPICAL-DYSARTHRIC SCENARIO

This subsection provides the full phoneme error rate results complementing the matched-size typical-dysarthric results presented in Section 5.2.1.

Table A.9 presents full results for monolingual experiments. Table A.10 presents full results for the bilingual CTC-only fine-tuning experiments. Table A.11 presents full results for the contrastive learning with basic sampling strategy experiments. Table A.12 presents full results for the contrastive learning with equivalence-prioritized strategy experiments. Table A.13 presents full results for the contrastive learning with hierarchical strategy experiments.

Table A.9: Phoneme error rate results for the matched-size typical-dysarthric scenario monolingual contrastive learning experiments across alignment types and triplet loss weights. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. The best result in each evaluation scope is underlined. Equiv. PER denotes PER computed over the language-specific phoneme equivalence set.

		English		Dutch	
Alignment	λ_{triplet}	PER	Equiv. PER	PER	Equiv. PER
CTC-only baseline		**<u>10.67</u>** S: 6.39 I: 2.03 D: 2.25	**15.35** S: 12.36 I: 1.77 D: 1.22	**15.22** S: 8.08 I: 2.66 D: 4.48	**<u>15.22</u>** S: 9.91 I: 1.55 D: 3.75
Non-blank	0.05	**11.02** S: 6.63 I: 1.91 D: 2.48	**16.98** S: 13.72 I: 1.49 D: 1.77	**15.33** S: 8.19 I: 2.73 D: 4.42	**17.16** S: 11.21 I: 1.98 D: 3.97
	0.10	**11.06** S: 6.49 I: 1.92 D: 2.65	**15.08** S: 12.23 I: 0.95 D: 1.90	**15.47** S: 8.08 I: 2.91 D: 4.48	**15.91** S: 10.09 I: 1.72 D: 4.09
	0.20	**11.04** S: 6.54 I: 1.98 D: 2.52	**<u>14.95</u>** S: 12.23 I: 1.36 D: 1.36	**15.23** S: 8.20 I: 2.72 D: 4.31	**15.99** S: 10.09 I: 2.24 D: 3.66
	0.30	**10.92** S: 6.61 I: 1.97 D: 2.33	**16.98** S: 12.91 I: 1.90 D: 2.17	**15.46** S: 8.17 I: 2.77 D: 4.52	**16.34** S: 10.30 I: 2.24 D: 3.79
	0.40	**11.23** S: 6.74 I: 1.94 D: 2.55	**16.98** S: 13.59 I: 1.36 D: 2.04	**15.28** S: 8.23 I: 2.70 D: 4.36	**15.86** S: 10.17 I: 1.85 D: 3.84
Contiguous	0.05	**10.96** S: 6.34 I: 2.17 D: 2.45	**16.17** S: 12.64 I: 1.63 D: 1.90	**15.24** S: 8.23 I: 2.60 D: 4.41	**16.03** S: 9.96 I: 1.90 D: 4.18
	0.10	**10.96** S: 6.34 I: 2.17 D: 2.45	**16.98** S: 13.72 I: 1.22 D: 2.04	**15.34** S: 8.17 I: 2.84 D: 4.33	**15.73** S: 10.26 I: 1.81 D: 3.66
	0.20	**10.96** S: 6.44 I: 2.01 D: 2.51	**16.85** S: 13.32 I: 1.22 D: 2.31	**15.18** S: 8.29 I: 2.63 D: 4.27	**15.91** S: 10.56 I: 1.72 D: 3.62
	0.30	**11.16** S: 6.50 I: 1.94 D: 2.71	**15.90** S: 11.96 I: 1.36 D: 2.58	**<u>15.12</u>** S: 8.27 I: 2.70 D: 4.15	**15.65** S: 10.22 I: 1.85 D: 3.58
	0.40	**11.07** S: 6.44 I: 2.03 D: 2.60	**16.03** S: 12.50 I: 1.36 D: 2.17	**15.44** S: 8.45 I: 2.74 D: 4.24	**15.91** S: 10.65 I: 1.81 D: 3.45

A

Table A.10: Phoneme error rate results for the matched-size typical-dysarthric scenario bilingual CTC-only fine-tuning. Substitution, insertion, and deletion rates are denoted by S, I, and D, respectively.

Evaluation scope	PER (%)	S (%)	I (%)	D (%)
Overall	13.21	7.66	2.09	3.46
Dutch	15.06	8.14	2.35	4.57
English	11.63	7.24	1.87	2.52
Equivalence phonemes	16.88	11.62	2.11	3.14
English equivalences	18.16	12.76	2.94	2.46
Dutch equivalences	16.08	10.91	1.59	3.58

Table A.11: Phoneme error rate (%) for the matched-size typical-dysarthric scenario basic CL across alignment types and triplet loss weights. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. The best available PER in each evaluation scope is underlined. Equiv. denotes evaluation over the phoneme equivalence set.

Alignment	λ_{triplet}	Overall	Dutch	English	Equiv.	EN equiv.	NL equiv.
Non-blank	0.05	**13.33**	**15.16**	**11.76**	**16.64**	**18.09**	**15.73**
		S: 7.85	S: 8.41	S: 7.38	S: 11.28	S: 13.04	S: 10.17
		I: 2.12	I: 2.40	I: 1.88	I: 2.17	I: 2.87	I: 1.72
		D: 3.35	D: 4.36	D: 2.50	D: 3.20	D: 2.18	D: 3.84
	0.10	**13.24**	**15.30**	**11.47**	**16.41**	**17.13**	**15.95**
		S: 7.84	S: 8.42	S: 7.34	S: 11.60	S: 12.83	S: 10.82
		I: 2.08	I: 2.31	I: 1.88	I: 1.88	I: 2.59	I: 1.42
		D: 3.32	D: 4.57	D: 2.25	D: 2.93	D: 1.71	D: 3.71
	0.20	**13.20**	**15.60**	**11.14**	**16.75**	**16.31**	**17.03**
		S: 7.68	S: 8.55	S: 6.93	S: 11.44	S: 11.95	S: 11.12
		I: 2.13	I: 2.47	I: 1.84	I: 2.17	I: 2.25	I: 2.11
		D: 3.39	D: 4.58	D: 2.37	D: 3.14	D: 2.12	D: 3.79
	0.30	**13.23**	**14.95**	**11.76**	**16.17**	**17.88**	**15.09**
		S: 7.81	S: 8.28	S: 7.42	S: 11.10	S: 12.90	S: 9.96
		I: 2.12	I: 2.23	I: 2.02	I: 2.11	I: 2.87	I: 1.64
		D: 3.30	D: 4.44	D: 2.32	D: 2.96	D: 2.12	D: 3.49
	0.40	**13.25**	**15.26**	**11.53**	**16.20**	**17.13**	**15.60**
		S: 7.74	S: 8.31	S: 7.26	S: 11.36	S: 12.76	S: 10.47
		I: 2.09	I: 2.42	I: 1.81	I: 1.85	I: 2.32	I: 1.55
		D: 3.41	D: 4.53	D: 2.45	D: 2.99	D: 2.05	D: 3.58
Contiguous	0.05	**13.23**	**15.40**	**11.38**	**16.25**	**16.38**	**16.16**
		S: 7.83	S: 8.64	S: 7.13	S: 11.41	S: 12.15	S: 10.95
		I: 2.18	I: 2.56	I: 1.85	I: 1.96	I: 2.12	I: 1.85
		D: 3.23	D: 4.20	D: 2.40	D: 2.88	D: 2.12	D: 3.36
	0.10	**13.28**	**15.30**	**11.56**	**16.17**	**16.72**	**15.82**
		S: 7.90	S: 8.58	S: 7.32	S: 11.47	S: 12.22	S: 10.99
		I: 2.15	I: 2.38	I: 1.95	I: 1.82	I: 2.32	I: 1.51
		D: 3.24	D: 4.35	D: 2.29	D: 2.88	D: 2.18	D: 3.32
	0.20	**13.37**	**15.38**	**11.66**	**16.41**	**17.20**	**15.91**
		S: 7.90	S: 8.47	S: 7.41	S: 11.73	S: 12.90	S: 10.99
		I: 2.11	I: 2.29	I: 1.95	I: 1.64	I: 2.18	I: 1.29
		D: 3.37	D: 4.62	D: 2.30	D: 3.04	D: 2.12	D: 3.62
	0.30	**13.23**	**15.56**	**11.24**	**16.91**	**17.47**	**16.55**
		S: 7.68	S: 8.51	S: 6.96	S: 11.55	S: 12.97	S: 10.65
		I: 2.08	I: 2.39	I: 1.82	I: 2.06	I: 2.25	I: 1.94
		D: 3.47	D: 4.66	D: 2.46	D: 3.30	D: 2.25	D: 3.97
	0.40	**13.26**	**15.04**	**11.73**	**16.12**	**17.41**	**15.30**
		S: 7.67	S: 8.00	S: 7.38	S: 10.89	S: 12.35	S: 9.96
		I: 2.23	I: 2.49	I: 2.01	I: 2.01	I: 2.66	I: 1.59
		D: 3.36	D: 4.55	D: 2.35	D: 3.22	D: 2.39	D: 3.75

A

Table A.12: Phoneme error rate (%) for the matched-size typical-dysarthric scenario weighted equivalence-prioritized setup across alignment types and triplet loss weights. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. The best available PER in each evaluation scope is underlined. Equiv. denotes evaluation over the phoneme equivalence set.

Alignment	λ_{triplet}	Overall	Dutch	English	Equiv.	EN equiv.	NL equiv.
Non-blank	0.05	**<u>13.08</u>**	**15.10**	**<u>11.36</u>**	**16.51**	**17.47**	**15.91**
		S: 7.66	S: 8.44	S: 6.99	S: 11.65	S: 12.83	S: 10.91
		I: 2.12	I: 2.32	I: 1.95	I: 2.17	I: 2.73	I: 1.81
		D: 3.30	D: 4.34	D: 2.41	D: 2.69	D: 1.91	D: 3.19
	0.10	**13.31**	**15.50**	**11.43**	**<u>15.98</u>**	**<u>16.18</u>**	**15.86**
		S: 7.67	S: 8.38	S: 7.07	S: 11.07	S: 12.01	S: 10.47
		I: 1.95	I: 2.28	I: 1.67	I: 1.77	I: 1.84	I: 1.72
		D: 3.68	D: 4.85	D: 2.69	D: 3.14	D: 2.32	D: 3.66
	0.20	**13.48**	**15.33**	**11.89**	**16.80**	**17.61**	**16.29**
		S: 7.92	S: 8.36	S: 7.54	S: 11.76	S: 12.97	S: 10.99
		I: 2.11	I: 2.37	I: 1.90	I: 1.90	I: 2.39	I: 1.59
		D: 3.45	D: 4.60	D: 2.46	D: 3.14	D: 2.25	D: 3.71
	0.30	**13.29**	**<u>14.88</u>**	**11.92**	**16.33**	**18.02**	**<u>15.26</u>**
		S: 7.87	S: 8.17	S: 7.61	S: 11.33	S: 13.52	S: 9.96
		I: 2.13	I: 2.35	I: 1.94	I: 1.98	I: 2.32	I: 1.77
		D: 3.29	D: 4.36	D: 2.38	D: 3.01	D: 2.18	D: 3.53
	0.40	**13.57**	**15.38**	**12.02**	**16.75**	**17.47**	**16.29**
		S: 7.85	S: 8.26	S: 7.50	S: 11.41	S: 12.15	S: 10.95
		I: 2.16	I: 2.55	I: 1.83	I: 2.14	I: 2.59	I: 1.85
		D: 3.55	D: 4.57	D: 2.68	D: 3.20	D: 2.73	D: 3.49
Contiguous	0.05	**13.49**	**15.42**	**11.85**	**16.78**	**18.02**	**15.99**
		S: 7.96	S: 8.30	S: 7.68	S: 11.78	S: 14.06	S: 10.34
		I: 2.15	I: 2.44	I: 1.90	I: 2.03	I: 2.46	I: 1.77
		D: 3.38	D: 4.68	D: 2.27	D: 2.96	D: 1.50	D: 3.88
	0.10	**13.44**	**15.50**	**11.68**	**16.49**	**17.95**	**15.56**
		S: 7.83	S: 8.33	S: 7.41	S: 11.20	S: 13.38	S: 9.83
		I: 2.18	I: 2.54	I: 1.88	I: 2.27	I: 2.66	I: 2.03
		D: 3.42	D: 4.63	D: 2.39	D: 3.01	D: 1.91	D: 3.71
	0.20	**13.37**	**15.29**	**11.73**	**16.43**	**17.54**	**15.73**
		S: 7.78	S: 8.29	S: 7.34	S: 11.23	S: 13.04	S: 10.09
		I: 2.14	I: 2.38	I: 1.94	I: 1.88	I: 2.18	I: 1.68
		D: 3.45	D: 4.62	D: 2.45	D: 3.33	D: 2.32	D: 3.97
	0.30	**13.41**	**15.40**	**11.71**	**16.64**	**16.86**	**16.51**
		S: 7.82	S: 8.26	S: 7.43	S: 11.76	S: 12.70	S: 11.16
		I: 2.16	I: 2.49	I: 1.87	I: 1.98	I: 2.18	I: 1.85
		D: 3.44	D: 4.65	D: 2.41	D: 2.91	D: 1.98	D: 3.49
	0.40	**13.23**	**15.32**	**11.44**	**16.59**	**17.00**	**16.34**
		S: 7.80	S: 8.50	S: 7.20	S: 11.78	S: 12.97	S: 11.03
		I: 2.24	I: 2.54	I: 1.98	I: 2.09	I: 2.25	I: 1.98
		D: 3.20	D: 4.29	D: 2.26	D: 2.72	D: 1.77	D: 3.32

Table A.13: Phoneme error rate (%) for the matched-size typical-dysarthric scenario unweighted hierarchical setup across alignment types and triplet loss weights. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. The best available PER in each evaluation scope is underlined. Equiv. denotes evaluation over the phoneme equivalence set.

Alignment	λ_{triplet}	Overall	Dutch	English	Equiv.	EN equiv.	NL equiv.
Non-blank	0.05	**13.38**	**15.41**	**11.65**	**17.15**	**17.68**	**16.81**
		S: 7.85	S: 8.40	S: 7.38	S: 12.07	S: 13.72	S: 11.03
		I: 2.17	I: 2.48	I: 1.90	I: 1.96	I: 2.32	I: 1.72
		D: 3.36	D: 4.54	D: 2.36	D: 3.12	D: 1.64	D: 4.05
	0.10	**13.19**	**14.84**	**11.78**	**16.25**	**17.54**	**15.43**
		S: 7.72	S: 8.07	S: 7.42	S: 11.33	S: 12.97	S: 10.30
		I: 2.09	I: 2.35	I: 1.87	I: 1.96	I: 2.53	I: 1.59
		D: 3.38	D: 4.42	D: 2.49	D: 2.96	D: 2.05	D: 3.53
	0.20	**13.51**	**15.51**	**11.79**	**16.86**	**18.02**	**16.12**
		S: 7.91	S: 8.52	S: 7.39	S: 11.57	S: 13.11	S: 10.60
		I: 2.29	I: 2.65	I: 1.99	I: 2.25	I: 3.00	I: 1.77
		D: 3.30	D: 4.34	D: 2.41	D: 3.04	D: 1.91	D: 3.75
	0.30	**13.38**	**15.46**	**11.60**	**16.75**	**17.13**	**16.51**
		S: 7.84	S: 8.49	S: 7.29	S: 11.86	S: 13.38	S: 10.91
		I: 2.23	I: 2.55	I: 1.95	I: 1.98	I: 2.18	I: 1.85
		D: 3.31	D: 4.42	D: 2.35	D: 2.91	D: 1.57	D: 3.75
	0.40	**13.19**	**15.22**	**11.46**	**16.04**	**16.45**	**15.78**
		S: 7.73	S: 8.23	S: 7.31	S: 10.99	S: 12.15	S: 10.26
		I: 2.16	I: 2.52	I: 1.85	I: 2.03	I: 2.32	I: 1.85
		D: 3.30	D: 4.48	D: 2.30	D: 3.01	D: 1.98	D: 3.66
Contiguous	0.05	**13.22**	**15.20**	**11.52**	**16.51**	**17.00**	**16.21**
		S: 7.78	S: 8.37	S: 7.28	S: 11.39	S: 12.22	S: 10.86
		I: 2.11	I: 2.33	I: 1.92	I: 1.93	I: 2.46	I: 1.59
		D: 3.33	D: 4.50	D: 2.32	D: 3.20	D: 2.32	D: 3.75
	0.10	**13.41**	**15.56**	**11.57**	**16.35**	**16.66**	**16.16**
		S: 7.72	S: 8.53	S: 7.03	S: 11.65	S: 12.63	S: 11.03
		I: 2.20	I: 2.38	I: 2.04	I: 1.82	I: 1.98	I: 1.72
		D: 3.49	D: 4.65	D: 2.50	D: 2.88	D: 2.05	D: 3.41
	0.20	**13.56**	**15.72**	**11.72**	**16.78**	**18.02**	**15.99**
		S: 7.79	S: 8.49	S: 7.20	S: 11.36	S: 12.90	S: 10.39
		I: 2.29	I: 2.70	I: 1.93	I: 2.22	I: 2.94	I: 1.77
		D: 3.48	D: 4.53	D: 2.59	D: 3.20	D: 2.18	D: 3.84
	0.30	**13.44**	**15.71**	**11.50**	**16.80**	**16.66**	**16.90**
		S: 7.89	S: 8.55	S: 7.31	S: 11.76	S: 12.35	S: 11.38
		I: 2.25	I: 2.56	I: 1.97	I: 2.19	I: 2.46	I: 2.03
		D: 3.31	D: 4.59	D: 2.22	D: 2.85	D: 1.84	D: 3.49
	0.40	**13.29**	**15.28**	**11.59**	**16.99**	**17.82**	**16.47**
		S: 7.82	S: 8.35	S: 7.36	S: 11.65	S: 12.97	S: 10.82
		I: 2.28	I: 2.59	I: 2.02	I: 2.35	I: 2.94	I: 1.98
		D: 3.19	D: 4.35	D: 2.20	D: 2.99	D: 1.91	D: 3.66

A.2.2. FULL DYSARTHRIC DATASET WITH ADDITIONAL TYPICAL DATA SCENARIO

This subsection provides the full phoneme error rate results complementing the additional typical data results presented in Section 5.2.2.

Table A.14 presents full results for monolingual experiments. Table A.15 presents full results for the bilingual CTC-only fine-tuning experiments. Table A.16 presents full results for the contrastive learning with basic sampling strategy experiments. Table A.17 presents full results for the contrastive learning with equivalence-prioritized strategy experiments. Table A.18 presents full results for the contrastive learning with hierarchical strategy experiments.

Table A.14: Phoneme error rate results for the additional typical data monolingual contrastive learning experiments across alignment types and triplet loss weights. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. The best result in each evaluation scope is underlined. Equiv. PER denotes PER computed over the language-specific phoneme equivalence set.

		English		Dutch	
Alignment	λ_{triplet}	PER	Equiv. PER	PER	Equiv. PER
CTC-only baseline		**7.58** S: 4.08 I: 1.53 D: 1.97	**11.96** S: 8.70 I: 1.49 D: 1.77	**10.14** S: 5.20 I: 1.77 D: 3.17	**10.22** S: 6.38 I: 1.29 D: 2.54
Non-blank	0.05	**7.84** S: 4.31 I: 1.44 D: 2.09	**11.82** S: 8.97 I: 1.36 D: 1.49	**9.98** S: 5.13 I: 1.97 D: 2.88	**9.87** S: 6.12 I: 1.42 D: 2.33
	0.10	**7.68** S: 4.18 I: 1.51 D: 1.99	**10.60** S: 7.74 I: 1.36 D: 1.49	**10.21** S: 5.19 I: 1.81 D: 3.21	**10.65** S: 6.77 I: 1.29 D: 2.59
	0.20	**7.96** S: 4.39 I: 1.37 D: 2.20	**12.64** S: 8.83 I: 1.22 D: 2.58	**10.18** S: 5.33 I: 1.98 D: 2.86	**10.47** S: 6.72 I: 1.77 D: 1.98
	0.30	**7.96** S: 4.39 I: 1.37 D: 2.20	**12.64** S: 8.83 I: 1.22 D: 2.58	**10.26** S: 5.27 I: 1.94 D: 3.05	**10.69** S: 6.85 I: 1.51 D: 2.33
	0.40	**7.88** S: 4.40 I: 1.53 D: 1.95	**11.28** S: 7.74 I: 1.36 D: 2.17	**10.06** S: 5.03 I: 1.89 D: 3.15	**10.39** S: 6.47 I: 1.51 D: 2.41
Contiguous	0.05	**7.77** S: 4.25 I: 1.53 D: 1.99	**12.09** S: 8.83 I: 1.22 D: 2.04	**10.06** S: 5.03 I: 1.89 D: 3.15	**10.39** S: 6.47 I: 1.51 D: 2.41
	0.10	**7.80** S: 4.30 I: 1.51 D: 1.99	**11.82** S: 8.15 I: 1.63 D: 2.04	**9.97** S: 5.23 I: 1.79 D: 2.94	**10.78** S: 6.90 I: 1.47 D: 2.41
	0.20	**7.84** S: 4.36 I: 1.50 D: 1.99	**11.96** S: 8.15 I: 1.63 D: 2.17	**9.77** S: 5.10 I: 1.85 D: 2.82	**9.87** S: 6.16 I: 1.38 D: 2.33
	0.30	**7.84** S: 4.22 I: 1.55 D: 2.08	**10.73** S: 7.47 I: 1.22 D: 2.04	**10.15** S: 5.31 I: 1.68 D: 3.15	**11.34** S: 6.94 I: 1.51 D: 2.89
	0.40	**7.99** S: 4.38 I: 1.67 D: 1.94	**12.09** S: 8.15 I: 1.77 D: 2.17	**10.50** S: 5.40 I: 1.80 D: 3.29	**11.08** S: 7.03 I: 1.16 D: 2.89

A

Table A.15: Phoneme error rate results for the additional typical data bilingual CTC-only fine-tuning. Substitution, insertion, and deletion rates are denoted by S, I, and D, respectively.

Evaluation scope	PER (%)	S (%)	I (%)	D (%)
Overall	8.67	4.90	1.47	2.30
Dutch	10.30	5.49	1.67	3.13
English	7.29	4.39	1.30	1.60
Equivalence phonemes	10.81	7.13	1.32	2.35
English equivalences	10.03	7.17	1.30	1.57
Dutch equivalences	11.29	7.11	1.34	2.84

Table A.16: Phoneme error rate (%) for the additional typical data scenario basic CL across alignment types and triplet loss weights. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. The best available PER in each evaluation scope is underlined. Equiv. denotes evaluation over the phoneme equivalence set.

Alignment	λ_{triplet}	Overall	Dutch	English	Equiv.	EN equiv.	NL equiv.
Non-blank	0.05	**8.70**	**10.27**	**7.37**	**10.67**	**10.24**	**10.95**
		S: 4.89	S: 5.47	S: 4.39	S: 7.32	S: 7.24	S: 7.37
		I: 1.59	I: 1.79	I: 1.42	I: 1.29	I: 1.50	I: 1.16
		D: 2.23	D: 3.01	D: 1.56	D: 2.06	D: 1.50	D: 2.41
	0.10	**8.71**	**10.24**	**7.41**	**11.04**	**10.38**	**11.47**
		S: 4.83	S: 5.34	S: 4.39	S: 7.29	S: 7.17	S: 7.37
		I: 1.55	I: 1.80	I: 1.33	I: 1.37	I: 1.43	I: 1.34
		D: 2.33	D: 3.09	D: 1.69	D: 2.38	D: 1.77	D: 2.76
	0.20	**8.83**	**10.25**	**7.61**	**10.89**	**10.65**	**11.03**
		S: 5.02	S: 5.65	S: 4.48	S: 7.53	S: 7.71	S: 7.41
		I: 1.51	I: 1.70	I: 1.35	I: 1.32	I: 1.37	I: 1.29
		D: 2.30	D: 2.90	D: 1.78	D: 2.03	D: 1.57	D: 2.33
	0.30	**8.78**	**10.18**	**7.58**	**10.89**	**10.85**	**10.91**
		S: 4.97	S: 5.51	S: 4.51	S: 7.37	S: 7.85	S: 7.07
		I: 1.53	I: 1.71	I: 1.37	I: 1.24	I: 1.37	I: 1.16
		D: 2.29	D: 2.97	D: 1.71	D: 2.27	D: 1.64	D: 2.67
	0.40	**8.74**	**10.34**	**7.37**	**10.73**	**10.17**	**11.08**
		S: 4.84	S: 5.47	S: 4.31	S: 6.74	S: 6.69	S: 6.77
		I: 1.43	I: 1.61	I: 1.28	I: 1.32	I: 1.57	I: 1.16
		D: 2.46	D: 3.26	D: 1.78	D: 2.67	D: 1.91	D: 3.15
Contiguous	0.05	**8.79**	**10.26**	**7.53**	**10.89**	**10.58**	**11.08**
		S: 4.98	S: 5.51	S: 4.52	S: 7.45	S: 7.24	S: 7.59
		I: 1.48	I: 1.70	I: 1.30	I: 1.48	I: 1.77	I: 1.29
		D: 2.33	D: 3.06	D: 1.71	D: 1.96	D: 1.57	D: 2.20
	0.10	**8.88**	**10.52**	**7.48**	**10.44**	**9.69**	**10.91**
		S: 4.92	S: 5.50	S: 4.43	S: 6.95	S: 6.62	S: 7.16
		I: 1.49	I: 1.70	I: 1.32	I: 1.37	I: 1.57	I: 1.25
		D: 2.47	D: 3.32	D: 1.74	D: 2.11	D: 1.50	D: 2.50
	0.20	**8.96**	**10.66**	**7.50**	**10.78**	**9.90**	**11.34**
		S: 4.96	S: 5.62	S: 4.39	S: 7.08	S: 6.83	S: 7.24
		I: 1.54	I: 1.79	I: 1.33	I: 1.45	I: 1.71	I: 1.29
		D: 2.46	D: 3.26	D: 1.78	D: 2.25	D: 1.37	D: 2.80
	0.30	**8.80**	**10.24**	**7.57**	**10.78**	**10.65**	**10.86**
		S: 5.01	S: 5.60	S: 4.50	S: 7.66	S: 7.58	S: 7.72
		I: 1.46	I: 1.68	I: 1.28	I: 1.14	I: 1.23	I: 1.08
		D: 2.33	D: 2.95	D: 1.80	D: 1.98	D: 1.84	D: 2.07
	0.40	**8.86**	**10.49**	**7.47**	**10.91**	**10.03**	**11.47**
		S: 4.96	S: 5.58	S: 4.43	S: 7.24	S: 6.69	S: 7.59
		I: 1.51	I: 1.70	I: 1.34	I: 1.51	I: 1.64	I: 1.42
		D: 2.39	D: 3.21	D: 1.70	D: 2.17	D: 1.71	D: 2.46

A

Table A.17: Phoneme error rate (%) for the additional typical data scenario weighted equivalence-prioritized setup across alignment types and triplet loss weights. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. The best available PER in each evaluation scope is underlined. Equiv. denotes evaluation over the phoneme equivalence set.

Alignment	λ_{triplet}	Overall	Dutch	English	Equiv.	EN equiv.	NL equiv.
Non-blank	0.05	**8.57**	**9.92**	**7.42**	**10.30**	**10.58**	**10.13**
		S: 4.90	S: 5.50	S: 4.39	S: 7.13	S: 7.78	S: 6.72
		I: 1.41	I: 1.55	I: 1.30	I: 1.27	I: 1.57	I: 1.08
		D: 2.26	D: 2.88	D: 1.73	D: 1.90	D: 1.23	D: 2.33
	0.10	**8.98**	**10.53**	**7.66**	**10.65**	**9.76**	**11.21**
		S: 4.99	S: 5.54	S: 4.52	S: 7.03	S: 6.69	S: 7.24
		I: 1.65	I: 1.94	I: 1.41	I: 1.72	I: 1.77	I: 1.68
		D: 2.33	D: 3.05	D: 1.72	D: 1.90	D: 1.30	D: 2.28
	0.20	**8.92**	**10.31**	**7.73**	**11.04**	**10.78**	**11.21**
		S: 5.00	S: 5.52	S: 4.56	S: 7.40	S: 7.51	S: 7.33
		I: 1.43	I: 1.60	I: 1.28	I: 1.32	I: 1.43	I: 1.25
		D: 2.49	D: 3.19	D: 1.89	D: 2.32	D: 1.84	D: 2.63
	0.30	**8.86**	**10.52**	**7.45**	**10.86**	**10.24**	**11.25**
		S: 4.97	S: 5.54	S: 4.48	S: 7.50	S: 7.44	S: 7.54
		I: 1.57	I: 1.81	I: 1.37	I: 1.11	I: 1.16	I: 1.08
		D: 2.32	D: 3.17	D: 1.60	D: 2.25	D: 1.64	D: 2.63
	0.40	**8.78**	**10.42**	**7.38**	**11.12**	**10.31**	**11.64**
		S: 4.95	S: 5.69	S: 4.32	S: 7.42	S: 7.30	S: 7.50
		I: 1.51	I: 1.70	I: 1.35	I: 1.40	I: 1.50	I: 1.34
		D: 2.32	D: 3.03	D: 1.71	D: 2.30	D: 1.50	D: 2.80
Contiguous	0.05	**8.69**	**10.24**	**7.36**	**10.36**	**9.62**	**10.82**
		S: 4.89	S: 5.50	S: 4.37	S: 7.16	S: 6.62	S: 7.50
		I: 1.46	I: 1.61	I: 1.32	I: 1.14	I: 1.43	I: 0.95
		D: 2.34	D: 3.12	D: 1.67	D: 2.06	D: 1.57	D: 2.37
	0.10	**8.79**	**10.20**	**7.59**	**10.59**	**10.03**	**10.95**
		S: 4.93	S: 5.39	S: 4.54	S: 7.08	S: 6.96	S: 7.16
		I: 1.53	I: 1.78	I: 1.32	I: 1.32	I: 1.37	I: 1.29
		D: 2.33	D: 3.03	D: 1.73	D: 2.19	D: 1.71	D: 2.50
	0.20	**8.78**	**10.04**	**7.70**	**10.44**	**10.72**	**10.26**
		S: 5.00	S: 5.46	S: 4.60	S: 7.45	S: 7.51	S: 7.41
		I: 1.62	I: 1.76	I: 1.50	I: 1.29	I: 1.71	I: 1.03
		D: 2.16	D: 2.82	D: 1.60	D: 1.69	D: 1.50	D: 1.81
	0.30	**8.95**	**10.57**	**7.57**	**11.15**	**10.85**	**11.34**
		S: 4.97	S: 5.60	S: 4.43	S: 7.40	S: 7.30	S: 7.46
		I: 1.54	I: 1.83	I: 1.30	I: 1.72	I: 1.98	I: 1.55
		D: 2.44	D: 3.15	D: 1.85	D: 2.03	D: 1.57	D: 2.33
	0.40	**8.58**	**10.09**	**7.29**	**10.91**	**10.78**	**10.99**
		S: 4.82	S: 5.45	S: 4.28	S: 7.34	S: 7.51	S: 7.24
		I: 1.53	I: 1.78	I: 1.32	I: 1.37	I: 1.71	I: 1.16
		D: 2.23	D: 2.85	D: 1.70	D: 2.19	D: 1.57	D: 2.59

Table A.18: Phoneme error rate (%) for the additional typical data scenario unweighted hierarchical setup across alignment types and triplet loss weights. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. The best available PER in each evaluation scope is underlined. Equiv. denotes evaluation over the phoneme equivalence set.

Alignment	λ_{triplet}	Overall	Dutch	English	Equiv.	EN equiv.	NL equiv.
Non-blank	0.05	**8.68**	**10.39**	**7.22**	**10.25**	**9.49**	**9.49**
		S: 4.84	S: 5.44	S: 4.32	S: 6.84	S: 6.76	S: 6.76
		I: 1.56	I: 1.82	I: 1.33	I: 1.35	I: 1.43	I: 1.43
		D: 2.28	D: 3.13	D: 1.56	D: 2.06	D: 1.30	D: 1.30
	0.10	**8.91**	**10.43**	**7.61**	**10.75**	**10.24**	**11.08**
		S: 4.93	S: 5.50	S: 4.44	S: 7.27	S: 7.30	S: 7.24
		I: 1.62	I: 1.91	I: 1.38	I: 1.19	I: 1.23	I: 1.16
		D: 2.36	D: 3.03	D: 1.79	D: 2.30	D: 1.71	D: 2.67
	0.20	**8.75**	**10.18**	**7.52**	**10.49**	**10.10**	**10.73**
		S: 4.75	S: 5.22	S: 4.35	S: 6.92	S: 7.51	S: 6.55
		I: 1.49	I: 1.72	I: 1.30	I: 1.29	I: 1.30	I: 1.29
		D: 2.50	D: 3.25	D: 1.86	D: 2.27	D: 1.30	D: 2.89
	0.30	**8.62**	**9.96**	**7.47**	**10.36**	**9.56**	**10.86**
		S: 4.79	S: 5.33	S: 4.32	S: 6.95	S: 6.76	S: 7.07
		I: 1.62	I: 1.79	I: 1.48	I: 1.45	I: 1.57	I: 1.38
		D: 2.21	D: 2.83	D: 1.67	D: 1.96	D: 1.23	D: 2.41
	0.40	**8.81**	**10.30**	**7.54**	**10.59**	**10.03**	**10.95**
		S: 4.90	S: 5.47	S: 4.42	S: 7.00	S: 6.69	S: 7.20
		I: 1.54	I: 1.78	I: 1.33	I: 1.24	I: 1.23	I: 1.25
		D: 2.37	D: 3.05	D: 1.79	D: 2.35	D: 2.12	D: 2.50
Contiguous	0.05	**8.64**	**10.06**	**7.43**	**10.65**	**9.22**	**11.55**
		S: 4.76	S: 5.29	S: 4.30	S: 6.90	S: 6.14	S: 7.37
		I: 1.59	I: 1.79	I: 1.43	I: 1.56	I: 1.64	I: 1.51
		D: 2.29	D: 2.98	D: 1.71	D: 2.19	D: 1.43	D: 2.67
	0.10	**8.73**	**10.12**	**7.54**	**10.36**	**10.44**	**10.30**
		S: 4.84	S: 5.30	S: 4.45	S: 6.84	S: 7.24	S: 6.59
		I: 1.54	I: 1.75	I: 1.35	I: 1.29	I: 1.30	I: 1.29
		D: 2.36	D: 3.07	D: 1.74	D: 2.22	D: 1.91	D: 2.41
	0.20	**8.89**	**10.33**	**7.66**	**11.10**	**10.44**	**11.51**
		S: 5.01	S: 5.68	S: 4.43	S: 7.53	S: 6.89	S: 7.93
		I: 1.53	I: 1.67	I: 1.41	I: 1.29	I: 1.57	I: 1.12
		D: 2.36	D: 2.98	D: 1.82	D: 2.27	D: 1.98	D: 2.46
	0.30	**8.61**	**9.94**	**7.46**	**10.09**	**9.90**	**10.22**
		S: 4.82	S: 5.23	S: 4.46	S: 6.82	S: 7.03	S: 6.68
		I: 1.50	I: 1.75	I: 1.30	I: 1.40	I: 1.30	I: 1.47
		D: 2.28	D: 2.97	D: 1.70	D: 1.88	D: 1.57	D: 2.07
	0.40	**8.69**	**10.23**	**7.38**	**10.59**	**10.58**	**10.60**
		S: 4.89	S: 5.43	S: 4.42	S: 7.16	S: 7.58	S: 6.90
		I: 1.47	I: 1.71	I: 1.26	I: 1.24	I: 1.37	I: 1.16
		D: 2.34	D: 3.09	D: 1.70	D: 2.19	D: 1.64	D: 2.54

A

A.3. RESULTS FOR EXPERIMENTS USING DYNAMIC ALIGNMENTS

This section provides the phoneme error rate results for the dynamic-alignment experiments described in Section 3.2.

Table A.19: Phoneme error rate (%) for dysarthric-data-only equivalence-prioritized strategy bilingual experiments using dynamic alignments. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. Equiv. denotes evaluation over the phoneme equivalence set.

Alignment	λ_{triplet}	Overall	Dutch	English	Equiv.	EN equiv.	NL equiv.
Non-blank	0.05	**10.35**	**11.82**	**9.09**	**12.02**	**12.83**	**11.51**
		S: 5.71	S: 6.16	S: 5.33	S: 8.06	S: 9.08	S: 7.41
		I: 1.80	I: 2.20	I: 1.46	I: 1.29	I: 1.71	I: 1.03
		D: 2.84	D: 3.47	D: 2.31	D: 2.67	D: 2.05	D: 3.06
	0.10	**10.99**	**12.50**	**9.71**	**12.76**	**13.24**	**12.46**
		S: 5.99	S: 6.46	S: 5.58	S: 8.45	S: 9.56	S: 7.76
		I: 1.97	I: 2.38	I: 1.61	I: 1.64	I: 1.98	I: 1.42
		D: 3.04	D: 3.65	D: 2.52	D: 2.67	D: 1.71	D: 3.28
	0.20	**10.78**	**12.34**	**9.45**	**12.68**	**13.04**	**12.46**
		S: 5.90	S: 6.43	S: 5.45	S: 8.56	S: 9.69	S: 7.84
		I: 1.81	I: 2.07	I: 1.60	I: 1.40	I: 1.71	I: 1.21
		D: 3.06	D: 3.84	D: 2.40	D: 2.72	D: 1.64	D: 3.41
	0.30	**10.20**	**11.84**	**8.79**	**12.39**	**12.56**	**12.28**
		S: 5.71	S: 6.31	S: 5.19	S: 8.38	S: 8.87	S: 8.06
		I: 1.72	I: 2.08	I: 1.42	I: 1.72	I: 1.98	I: 1.55
		D: 2.77	D: 3.45	D: 2.19	D: 2.30	D: 1.71	D: 2.67

Table A.20: Phoneme error rate (%) for dysarthric-data-only hierarchical strategy bilingual experiments using dynamic alignments. Each cell reports PER in bold, followed by substitution (S), insertion (I), and deletion (D) rates. Equiv. denotes evaluation over the phoneme equivalence set.

Alignment	λ_{triplet}	Overall	Dutch	English	Equiv.	EN equiv.	NL equiv.
Non-blank	0.05	**10.45**	**12.11**	**9.03**	**12.60**	**11.95**	**13.02**
		S: 5.77	S: 6.28	S: 5.33	S: 8.19	S: 8.40	S: 8.06
		I: 1.82	I: 2.22	I: 1.48	I: 1.72	I: 1.64	I: 1.77
		D: 2.86	D: 3.62	D: 2.22	D: 2.69	D: 1.91	D: 3.19
	0.10	**10.83**	**12.35**	**9.54**	**12.68**	**13.04**	**12.46**
		S: 5.97	S: 6.44	S: 5.56	S: 8.67	S: 9.22	S: 8.32
		I: 1.97	I: 2.35	I: 1.65	I: 1.74	I: 2.18	I: 1.47
		D: 2.89	D: 3.56	D: 2.32	D: 2.27	D: 1.64	D: 2.67
	0.20	**11.05**	**12.84**	**9.53**	**12.76**	**12.22**	**13.10**
		S: 6.13	S: 6.83	S: 5.52	S: 8.59	S: 8.53	S: 8.62
		I: 1.96	I: 2.32	I: 1.65	I: 1.69	I: 1.98	I: 1.51
		D: 2.96	D: 3.68	D: 2.35	D: 2.48	D: 1.71	D: 2.97
	0.30	**10.64**	**12.26**	**9.25**	**11.86**	**12.29**	**11.59**
		S: 5.96	S: 6.61	S: 5.40	S: 8.16	S: 8.81	S: 7.76
		I: 1.85	I: 2.20	I: 1.55	I: 1.35	I: 1.64	I: 1.16
		D: 2.83	D: 3.44	D: 2.30	D: 2.35	D: 1.84	D: 2.67