The representation of speech and its processing in the human brain and deep neural networks

Scharenborg, Odette

DOI
10.1007/978-3-030-26061-3_1

Publication date
2019

Document Version
Final published version

Published in
Speech and Computer

Citation (APA)
Scharenborg, O. (2019). The representation of speech and its processing in the human brain and deep neural networks. In A. A. Salah, A. Karpov, & R. Potapova (Eds.), *Speech and Computer: 21st International Conference, SPECOM 2019, Proceedings* (pp. 1-8). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 11658 LNAI). Springer. https://doi.org/10.1007/978-3-030-26061-3_1

The Representation of Speech and Its Processing in the Human Brain and Deep Neural Networks

Odette Scharenborg(✉)

Multimedia Analysis Group, Delft University of Technology,
Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands
o.e.scharenborg@tudelft.nl

Abstract. For most languages in the world and for speech that deviates from the standard pronunciation, not enough (annotated) speech data is available to train an automatic speech recognition (ASR) system. Moreover, human intervention is needed to adapt an ASR system to a new language or type of speech. Human listeners, on the other hand, are able to quickly adapt to nonstandard speech and can learn the sound categories of a new language without having been explicitly taught to do so. In this paper, I will present comparisons between human speech processing and deep neural network (DNN)-based ASR and will argue that the cross-fertilisation of the two research fields can provide valuable information for the development of ASR systems that can flexibly adapt to any type of speech in any language. Specifically, I present results of several experiments carried out on both human listeners and DNN-based ASR systems on the representation of speech and lexically-guided perceptual learning, i.e., the ability to adapt a sound category on the basis of new incoming information resulting in improved processing of subsequent speech. The results showed that DNNs appear to learn structures that humans use to process speech without being explicitly trained to do so, and that, similar to humans, DNN systems learn speaker-adapted phone category boundaries from a few labelled examples. These results are the first steps towards building human-speech processing inspired ASR systems that, similar to human listeners, can adjust flexibly and fast to all kinds of new speech.

Keywords: Speech representations · Adaptation · Non-standard speech · Deep neural networks · Human speech processing · Perceptual learning

1 Introduction

Automatic speech recognition (ASR) is the mapping of a continuous, highly variable speech signal onto discrete, abstract representations, typically phonemes or words. ASR works well in relatively restricted settings (e.g., speech without strong accents, quiet background) but tends to break down when the speech that needs to be recognised diverges from 'normal' speech, e.g., because of speech impediments or accents, or when no or only limited annotated training data is available for the type of speech or language for which the system is build (i.e., low-resource languages). In fact, for only

A. A. Salah et al. (Eds.): SPECOM 2019, LNAI 11658, pp. 1–8, 2019.
https://doi.org/10.1007/978-3-030-26061-3_1

about 1% of the world languages the minimum amount of training data that is needed to develop ASR technology is available [1]. This means that ASR technology is not available to all people in the world, including those people who would profit the most from it, i.e., people with disabilities or people whose native language does not have a common written form, because of which they need to rely on speech technology to communicate with people and/or computers. The Linguistic Rights as included in the Universal Declaration of Human Rights states that it is a human right to communicate in one's native language. This situation is obviously not yet reached.

Most ASR systems are phoneme-based systems which are based on the principle that a word is composed of a sequence of speech sounds called phonemes, and acoustic representations (i.e., acoustic models) are trained for context-dependent versions of each phoneme. In order to make ASR available for all types of speech in all the world's languages, simply recording and annotating enough speech material for training a phoneme-based ASR system is infeasible. First, because it is impossible to collect for every type of speech in every language of the world the hundreds of hours of speech with their textual transcriptions that are needed to train a system that works reasonably well. Second, because it is impossible for languages that do not have a common writing system. An obvious solution is to map an ASR system trained on a language and type of speech for which there is enough training data (e.g., English spoken by native speaker without a clear accent or speech impediment) to a language or type of speech for which there is little or no data [2–6]. This mapping of an ASR system from one type of speech or language to another requires explicit decisions by a human about which phoneme categories, or rather acoustic models, will need to be adapted or created in the ASR system.

In order to build ASR systems that can flexibly adapt to any type of speech in any language, we need: (1) invariant units of speech which transfer easily and accurately to other languages and different types of speech and lead to the best ASR recognition performance; (2) an ASR system that can flexibly adapt to new types of speech; (3) an ASR system that can decide when to create a new phoneme category, and do so.

Human listeners have been found to do exactly that. They are able to quickly adapt their phoneme categories on the basis of only a few examples to deviant speech, whether due to a speech impediment or an accent, using a process called lexically-guided perceptual learning [7]. Moreover, human listeners have been found to create new phoneme categories, e.g., when learning a new language [8].

So ideally, the search for invariant speech units and flexible adaptation processes in ASR are based on the speech representations and speech recognition processes in human speech processing as the best speech recogniser is a human who is a native speaker of the language [9]. Moreover, despite the differences in hardware between a human listener and an ASR system, they both carry out the same process: the recognition of speech [10]. There is ample evidence that knowledge about human speech processing has powerful potential for improving ASR ([9–14], for a review [15]). For instance, knowledge about human speech processing and human hearing has been used in the development of Mel-frequency cepstral coefficients (MFCCs, [16]) and Perceptual Linear Predictives (PLPs, [14]), while the episodic theory of human speech processing was the inspiration to the development of template-based approaches to ASR (e.g., [17]).

In this paper, I focus on the first two requisites for building an ASR system that can flexibly adapt to any type of speech in any language by comparing human speech processing and deep neural network (DNN)-based ASR. I will summarise experiments which compare the representation of speech and adaptation processes in the human brain and DNNs, with the ultimate aim to build human speech processing inspired ASR systems that can flexibly adapt to any speech style in any language, i.e., the third requisite. Recent advances in deep learning make DNNs currently the best-performing ASR systems [18]. DNNs are inspired by the human brain, which is often suggested to be the reason for their impressive abilities, e.g., [19]. Although both the human brain and DNNs consist of neurons and neural connections, little is known about whether DNNs actually use similar representations for speech and solve the task of speech recognition in the same way the human brain does.

Fig. 1. What features does a DNN use to distinguish between the plane in the blue sky on the left and the chair on the green lawn?

2 Speech Representations

When learning one's first language, human listeners learn to associate certain acoustic variability with certain phonological categories. The question I am interested in is whether a DNN also learns phonological categories similar to those used by human listeners. Using the visual example in Fig. 1 as an example: if a DNN is able to distinguish between a plane in a blue sky and a chair on a green lawn, has the DNN learned to distinguish the blue background from the green background or has it learned features that are associated with planes and features that are associated with chairs to distinguish the two objects as a human would do to distinguish these two objects in these pictures?

The question what speech representations a DNN learns during speech processing was investigated using a naïve, general feed-forward DNN which was trained on the task of vowel/consonant classification [20]. Vowel/consonant classification is a relatively simple, well-understood task, which allows us to investigate what a naïve, general DNN exactly learns when faced with the large variability of the speech sounds in the speech signal. Crucially, the speech representations in the different hidden layers of the DNN were investigated by visualising the activations of the speech representations in those hidden layers using different linguistic labels that are known to correspond to the underlying structures that human listeners use to process and understand speech.

The DNN consisted of 3 hidden layers with 1024 nodes each, and was trained on 64 h of read speech from the Corpus Spoken Dutch (CGN; [21]). Accuracy on the vowel/consonant classification task, averaged over five runs, was 85.5% (consonants: 85.2%; vowels: 86.7% correct). Subsequently, the input frames were labelled with:

- Phoneme labels: 39 in total.
- Manner of articulation: indicates the type of constriction in the vocal tract. For consonants, four categories were distinguished: plosive, fricative, nasal, approximant. For vowels, three categories were distinguished: short vowel, long vowel, diphthong.
- Place of articulation: indicates the location of the constriction in the vocal tract. For consonants, six categories were distinguished: bilabial, labiodental, velar, alveolar, palatal, glottal. For vowels, three tongue position categories were distinguished: front, central, back.

The clusters of speech representations at the different hidden layers were visualised using t-distributed neighbor embedding (t-SNE, [22]). The first visualisation investigated the clusters of consonants and vowels in the different hidden layers. The results showed that from earlier to later hidden layers, the vowel and consonant clusters become more compact and more separate, showing that the DNN is learning to create speech representations that are increasingly abstract.

In the second series of visualisations, the input frames were first labelled with the phoneme labels. This visualisation showed that the phoneme labels were not randomly distributed over the hidden layers. Rather, despite that the DNN was trained on a vowel/consonant classification task, the DNN implicitly learned to cluster frames with the same phoneme label to some extent. Subsequent analyses with labelling of the frames in terms of manner of articulation and place of articulation showed that the DNN learned to cluster sounds together that are produced in similar ways such that consonants with a similar manner of articulation and vowels with a similar place of articulation are clustered into clearly defined groups. The DNN thus appeared to learn structures that human listeners use to process speech without having been explicitly trained to do so.

3 Adaptation to Non-standard Speech

Adaptation to nonstandard speech is often referred to as ‘perceptual learning’ in the human speech processing literature. Perceptual learning is defined as the temporary or more permanent adaptation of sound categories after exposure to nonstandard speech such that the nonstandard sound is included into a pre-existing sound category, which leads to an improvement in the intelligibility of the speech (see for a review [23]). Perceptual learning is fast. Human listeners need only a few instances of the deviant sounds [24, 25] to adapt their sound category boundaries to include the nonstandard sound [7, 23–28]. ASR systems adapt to new speakers and listening conditions using both short-time adaptation algorithms (e.g., fMLLR [29]) and longer-term adaptation techniques (e.g., DNN weight training [30]). For both human listeners and ASR

systems, lexical knowledge about the word in which the nonstandard sound occurs is crucial to correctly interpret the nonstandard sound [7, 23].

3.1 Does a DNN Show Human-Like Adaptation to Nonstandard Speech?

In recent work, we investigated the question whether DNNs are able to adapt to nonstandard speech as rapidly as human listeners, and whether DNNs use intermediate speech representations that correlate with those used in human perceptual learning [12]. Mimicking the set-up of a human lexically-guided perceptual learning study [28], which allows for the direct comparison between human listening behaviour and the behaviour of the DNN, we trained a feed-forward DNN on the read speech of CGN. The trained model was regarded as a 'native Dutch listener'. In the next step, the DNN was retrained with the acoustic stimuli from the original human perceptual learning study [28], i.e., speech from a new speaker who had an (artificially created) nonstandard pronunciation of a sound in between [l] and [ɹ], referred to as [l/ɹ]: One model was trained with the [l/ɹ] sound always occurring in /r/-final words; another model was trained with the [l/ɹ] sound always occurring in /l/-final words. A final, baseline model was trained on the same words but without nonstandard pronunciations.

The results showed that the DNNs retrained with the [l/ɹ] sounds indeed showed perceptual learning: The baseline model classified the nonstandard sound during a subsequent phase as both [l] and [ɹ], the model retrained with the [l/ɹ] sound in /r/-final words classified the sound as [ɹ] while the model retrained with the nonstandard sound in /l/-final words classified the sound as [l]. This difference between the two models trained with the nonstandard pronunciation is called the perceptual learning effect. Moreover, this perceptual learning effect did not only occur at the output level, but calculations of the distances between the average activations of the nonstandard sound and those of the natural sounds and the visualisations of the activations of the hidden layers showed that perceptual learning also occurred at the DNN's intermediate levels. Interestingly, the visualisations of the speech representations in the DNN's hidden layers showed that the phonetic space was warped to accommodate the nonstandard speech. This warping of the phonetic space seems to be at odds with theories of human speech processing, which assumes that the nonstandard sound is incorporated in the existing phoneme category by redrawing the phoneme category boundaries [26]. In follow-up research, I plan to test this prediction of the DNN about human speech processing in new human perceptual experiments.

3.2 Are Nonstandard Sounds Processed Similarly in Human Listeners and DNNs?

In subsequent work, this research was pushed further and we asked the questions whether nonstandard sounds are processed in the same way as natural sounds; and, how many examples of the nonstandard sound are needed before the DNN adapts? Again, the experimental design [24] and acoustic stimuli were taken from earlier research on lexically-guided perceptual learning in human listeners [28]. The same DNN as in the study described in Sect. 3.1, was retrained but this time using increasing amounts of nonstandard sounds (in 10 bins of 4 ambiguous items). Calculations of the distances

between the average activations of the nonstandard sound and those of the natural sounds in the different hidden layers showed that the DNN showed perceptual learning after only four examples of the nonstandard sound, and little further adaptation for subsequent training examples.

Interestingly, human listeners have been found to show a similar type of step-like function after about 10–15 examples of the nonstandard sound. The difference in number of examples could be explained by the fact that the DNN sees each training example 30 times (30 epochs) whereas the human listener hears each token only once. In follow-up research, I plan to further investigate the step-like function in adaptation in human listening.

4 Concluding Remarks

In this paper, I summarised results from three studies comparing human speech processing and speech processing in deep neural networks. The results showed that:

- Similar to human listeners, the DNN progressively abstracted away variability in the speech signal in subsequent hidden layers;
- Without being explicitly trained to do so, the DNN captured the structure in speech by clustering the speech signal into linguistically-defined speech category representations, similar to those used during human speech processing;
- Similar to human listeners, the DNN adapted to nonstandard speech on the basis of only a few labelled examples by warping the phoneme space;
- This adaptation did not only occur in the output layer but instead occurred in the hidden layers of the DNN and showed a step-like function.

These detailed comparisons between human speech processing and DNN-based ASR highlight clear similarities between the speech representations and their processing in the human brain and in DNN-based ASR systems. Moreover, the DNNs made specific predictions about adaptation to nonstandard speech that will be investigated in experiments on human speech processing to further investigate the differences and similarities between adaptation to nonstandard speech in humans and DNN-based ASR systems. These experiments will lead to important new insights regarding the adaptation of human listeners to nonstandard speech.

Past research [9–17] has shown that knowledge of human speech processing can be used to improve ASRs. The observed similarities between human and DNN speech processing suggest that integrating the flexibility of the human adaptation processes into DNN-based ASRs is likely to lead to improved adaptation of DNN-based ASRs to nonstandard speech. Crucial for the development of human-speech processing inspired ASR systems that, similar to human listeners, can adjust flexibly and fast to all types of speech in all languages is understanding when and how human listeners decide to create a new phoneme category rather than adapting an existing phoneme category to include a nonstandard pronunciation. This is a crucial next step in this research.

Acknowledgments. I would like to thank Junrui Ni for carrying out the experiments described in Sect. 3.2 and Mark Hasegawa-Johnson for fruitful discussions on the experiments in Sect. 3.2.

References

1. Adda, G., et al.: Breaking the unwritten language barrier: the BULB project. In: Proceedings 5th Workshop on Spoken Language Technologies for Under-Resourced Languages (2016)
2. Waibel, A., Schultz, T.: Experiments on cross-language acoustic modelling. In: Proceedings of Interspeech (2001)
3. Vu, N.T., Metze, F., Schultz, T.: Multilingual bottleneck features and its application for under-resourced languages. In: Proceedings of the 3rd Workshop on Spoken Language Technologies for Under-Resourced Languages, Cape Town, South Africa (2012)
4. Xu, H., Do, V.H., Xiao, X., Chng, E.S.: A comparative study of BNF and DNN multilingual training on cross-lingual low-resource speech recognition. In: Proceedings of Interspeech, pp. 2132–2136 (2015)
5. Scharenborg, O., Ebel, P., Ciannella, F., Hasegawa-Johnson, M., Dehak, N.: Building an ASR system for Mboshi using a cross-language definition of acoustic units approach. In: Proceedings of the International Workshop on Spoken Language Technologies for Under-Resourced Languages, Gurugram, India (2018)
6. Scharenborg, O., et al.: Building an ASR system for a low-resource language through the adaptation of a high-resource language ASR system: preliminary results. In: Proceedings of the International Conference on Natural Language, Signal and Speech Processing, Casablanca, Morocco (2017)
7. Norris, D., McQueen, J.M., Cutler, A.: Perceptual learning in speech. Cogn. Psychol. **47**(2), 204–238 (2003)
8. Best, C.T., Tyler, M.C.: Nonnative and second-language speech perception. Commonalities and complementaries. In: Bohn, O.-S., Munro, M.J. (eds.) Language Experience in Second Language Speech Learning: In Honor of James Emil Flege, pp. 13–34. John Benjamins, Amsterdam (2007)
9. Davis, M.H., Scharenborg, O.: Speech perception by humans and machines. In: Gaskell, M.G., Mirkovic, J. (eds.) Speech Perception and Spoken Word Recognition, Part of the Series "Current Issues in the Psychology of Language", pp. 181–203. Routledge, London (2017)
10. Scharenborg, O., Norris, D., ten Bosch, L., McQueen, J.M.: How should a speech recognizer work? Cogn. Sci. **29**(6), 867–918 (2005)
11. Scharenborg, O.: Modeling the use of durational information in human spoken-word recognition. J. Acoust. Soc. Am. **127**(6), 3758–3770 (2010)
12. Scharenborg, O., Tiesmeyer, S., Hasegawa-Johnson, M., Dehak, N.: Visualizing phoneme category adaptation in deep neural networks. In: Proceedings of Interspeech (2018)
13. Dusan, S., Rabiner, L.R.: On integrating insights from human speech recognition into automatic speech recognition. In: Proceedings of Interspeech, pp. 1233–1236 (2005)
14. Hermansky, H.: Should recognizers have ears? Speech Commun. **25**, 3–27 (1998)
15. Scharenborg, O.: Reaching over the gap: a review of efforts to link human and automatic speech recognition research. Speech Commun. **49**, 336–347 (2007)
16. Davis, S., Mermelstein, P.: Comparison of the parametric representation for monosyllabic word recognition. IEEE Trans. Acoust. Speech Signal Process. **28**(4), 357–366 (1980)
17. De Wachter, M., Demuynck, K., van Compernolle, D., Wambaq, P.: Data driven example based continuous speech recognition. In: Proceedings of Eurospeech, Geneva, Switzerland, pp. 1133–1136 (2003)
18. Hinton, G., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Sig. Process. Mag. **29**(6), 82–97 (2012)

19. Wan, J., et al.: Deep learning for content-based image retrieval: a comprehensive study. In: Proceedings of the 22nd ACM International conference on Multimedia (MM 2014), pp. 157–166 (2014)
20. Scharenborg, O., van der Gouw, N., Larson, M., Marchiori, E.: The representation of speech in deep neural networks. In: Kompatsiaris, I., Huet, B., Mezaris, V., Gurrin, C., Cheng, W.-H., Vrochidis, S. (eds.) MMM 2019. LNCS, vol. 11296, pp. 194–205. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-05716-9_16
21. Oostdijk, N.H.J., et al.: Experiences from the spoken Dutch Corpus project. In: Proceedings of LREC, pp. 340–347 (2002)
22. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)
23. Samuel, A.G., Kraljic, T.: Perceptual learning in speech perception. Atten. Percept. Psychophys. **71**, 1207–1218 (2009)
24. Drozdova, P., van Hout, R., Scharenborg, O.: Processing and adaptation to ambiguous sounds during the course of perceptual learning. In: Proceedings of Interspeech, pp. 2811–2815 (2016)
25. Poellmann, K., McQueen, J.M., Mitterer, H.: The time course of perceptual learning. In: Proceedings of ICPhS (2011)
26. Clarke-Davidson, C., Luce, P.A., Sawusch, J.R.: Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? Percept. Psychophys. **70**, 604–618 (2008)
27. Drozdova, P., van Hout, R., Scharenborg, O.: Lexically-guided perceptual learning in non-native listening. Bilingualism: Lang. Cogn. **19**(5), 914–920 (2016). https://doi.org/10.1017/s136672891600002x
28. Scharenborg, O., Janse, E.: Comparing lexically-guided perceptual learning in younger and older listeners. Atten. Percept. Psychophys. **75**(3), 525–536 (2013). https://doi.org/10.3758/s13414-013-0422-4
29. Gales, M.J.: Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. **12**(2), 75–98 (1998)
30. Liao, H.: Speaker adaptation of context dependent deep neural networks. In: Proceedings of ICASSP, pp. 7947–7951 (2013)