Methods for Inferring the Phone Set of an Unwritten Language

Hasegawa-Johnson, Mark; Chen, Wenda; Scharenborg, Odette

Publication date
2018

Document Version
Accepted author manuscript

Citation (APA)
Hasegawa-Johnson, M., Chen, W., & Scharenborg, O. (2018). *Methods for Inferring the Phone Set of an Unwritten Language*. Abstract from 6th International Conference on Statistical Language and Speech Processing, Mons, Belgium.

Methods for Inferring the Phone Set of an Unwritten Language

Mark Hasegawa-Johnson, Wenda Chen, and Odette Scharenborg

Poster Submission to SLSP 2018

In engineering applications, phones are the representation intermediate between text and speech in many text-to-speech (TTS) and speech-to-text (STT) systems. When a language has no written form, TTS and STT are no longer meaningful acronyms as there is no text; we consider instead XTS and STX, where X is some other representation that can be easily interpreted by a human user, for example, image, translation, or chat. This paper presents experimental results from two speech applications for unwritten languages: image-to-speech, and speech-to-chat. Experimental evidence from these two applications suggests that the performance of an XTS or STX application can be significantly improved by defining or inferring a phone set for the unwritten language.

Image-to-speech (ITS) is the task of generating a spoken description of an image, in a language that has no written form. ITS can be trained and tested as a neural sequence-to-sequence transduction problem, in which an input sequence of sub-images is encoded, attended, and converted into a sequence of phone symbols, from which an output audio signal can be generated. The quality of ITS output varies dramatically depending on the quality of the phone set. Cheating experiments using a known correct phone set resulted in intelligible and meaningful spoken descriptions, but experiments using a cross-language phone set, or one automatically created using unsupervised methods, do not. Extrapolating beyond current experimental results, a simulated annealing algorithm will be presented that may be capable of finding the globally optimal phone set for matching a given ITS training database.

Speech-to-chat (STC) is the task of converting speech into a variably spelled transcription in the Latin alphabet, similar to the Latin-alphabet transcriptions used in online chat forums to represent colloquial dialects of multi-register languages such as Arabic and Hindi. Such chat transcripts can be easily collected, even from non-speakers of the language. When a non-speaker of the language writes down what she hears using a chat alphabet, she tends to map every phoneme in the utterance language to the most similar phoneme in her own language, where similarity can be defined by a weighted L1 distance between articulatory feature vectors. For this reason, the speech-to-chat paradigm allows us to infer a phone set that's actually pretty close to the unknown phoneme set of the unwritten language. Experiments were performed in which pseudo-under-resourced languages (Cantonese and Vietnamese, neither of which is truly "unwritten," though few people know how to write Cantonese) were transcribed by native speakers, and phonemic transcripts were generated from their transcriptions. Chat-alphabet transcriptions by non-speakers of Cantonese were then clustered in order to estimate the phonemic transcript. Extra information about the Cantonese phonemes (e.g., elicited from non-native transcribers with more than one native language) improves the quality of transcription.

We interpret these two results to mean that defining a better phone set for an unwritten language improves the quality of both image-to-speech and speech-to-chat applications.