Word Characters and Phone Pronunciation Embedding for ASR Confidence Classifier
Kshitiz Kumar, Tasos Anastasakos, Yifan Gong
{Kshitiz.Kumar, Tasos.Anastasakos, Yifan.Gong}@Microsoft.com

Scope and Abstract
- Confidences are integral to ASR systems, and applied to data selection, adaptation, ranking hypotheses, arbitration etc.
- Hybrid ASR system is inherently a match between pronunciations and AM+LM evidence but current confidence features lack pronunciation information.
- We develop pronunciation embeddings to represent and factorize acoustic score in relevant bases, and demonstrate 8-10% relative reduction in false alarm (FA) on large scale tasks.
- We generalize to standard NLP embeddings like Glove, and show 16% relative reduction in FA in combination with Glove.

Confidence Classifier Features and Training
- Acoustic (typically most important) and language model scores
- Background, Silence and Noise model scores
- Fanout, Perplexity, Duration features etc.
- We use MLP or deep learning models for confidence models

Motivation for Word Pronunciation Embeddings
- Acoustic score is aggregated over underlying model states (senones), and equivalently over word pronunciations.
- Acoustic scores differ for speakers and acoustic scenarios but also vary across correctly recognized words.
- We develop pronunciation embeddings as bases for the confidence model to learn score factorization over words.

Word Characters/Letters Embeddings
- For en-US, we use the 26 letters as bases and use letter count as features
- Ex – “cortana” has embedding in (2,1,1,1,1,1) at locations for {a,c,n,o,r,t}

Key Benefits of Proposed Embeddings
- Model learns acoustic score dependencies over pronunciation bases
- Smaller dimensional features (26-40), easy to train
- Relevant for low resource device settings, easily computed at runtime without extra storage requirements in Glove
- Incorporates elements of word identity
- Simple extension to other languages

Word Pronunciation Embeddings
- We use lexicon and create embeddings from monophone counts
- Closer to ASR search, retains most benefits of letter embeddings
- Avoids identical letter embeddings for word anagrams
  - Ex – Polo vs. Pool ("p ow l ow" vs. "p uw l f")
- Incorporates multiple pronunciations across dialects, accents

Experiments
- Applied to large scale enUS Server tasks across Mobile/Desktop/Bing
- Also applied to limited vocabulary Xbox tasks
- Xbox OOG task consists of movie or meetings
- MSE, CA and FA metrics

Higher Ranked Features constitute vowels

Experiments
- Applied to large scale enUS Server tasks across Mobile/Desktop/Bing
- Also applied to limited vocabulary Xbox tasks
- Xbox OOG task consists of movie or meetings
- MSE, CA and FA metrics

Higher Ranked Features constitute vowels

Conclusion
- Developed pronunciation embeddings to implicitly factorize acoustic score.
- Extended and combined the proposed embeddings with NLP embeddings like Glove/FastText.
- Letter embeddings showed 8.7% relative reduction in FA for Server task at CA~75%, along with 16.1% in combination with Glove.
- We can consider newer embeddings like BERT and also develop the embeddings work for ASR.