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

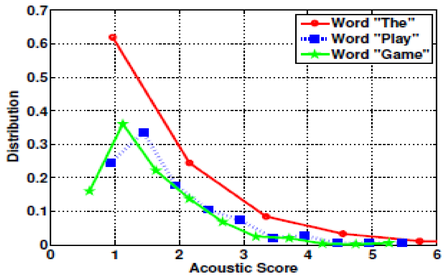
- Confidences are integral to ASR systems, and applied to data selection, adaptation, ranking hypotheses, arbitration etc.
- Hybrid ASR system is inherently a match between pronunciations and AM+LM evidence but current confidence features lack pronunciation information.
- We develop pronunciation embeddings 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.



- Embeddings normalize confidence scores for a global confidence threshold across 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}

Embedding Types	Representation for "cortana"
Character/Letter	c o r t a n a
Pronunciation	k ao r t ae n ax k ao r t aa n aa

Key Benefits of Proposed Embeddings

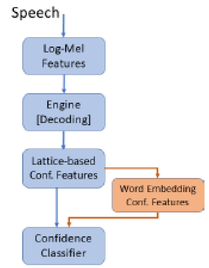
- Model learns acoustic score dependencies over pronunciation bases
- Smaller dimensional features (26-40), easy to train
- Relevant for low resource device settings, easily computed at runtime without extra storage requirements in Glove
- Incorporates elements of word identity
- Simple extension to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

Embeddings	Higher ranked features (in order)
Character/Letter	u, o, i, e, a
Pronunciation	eh, ey, iy, ay, ax

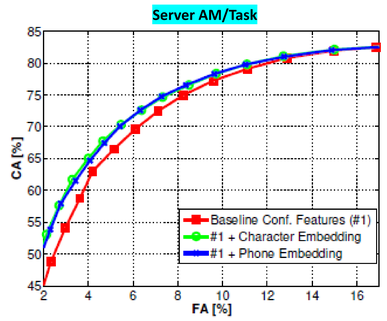
MSE Criterion

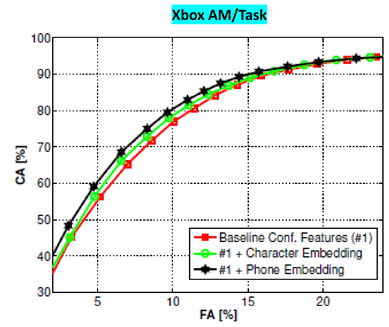
Items	Confidence Features	Validation MSE
1	Letter embedding	0.221
2	Acoustic Conf. feature	0.216
3	(2) + (1)	0.199
4	All Conf. features	0.188
5	(4) + (1)	0.183

Items	Confidence Features	Validation MSE
1	Pronunciation embedding	0.213
2	Acoustic conf. feature + (1)	0.195
3	All Conf. features + (1)	0.175

CA-FA Results for Server and Xbox AMs

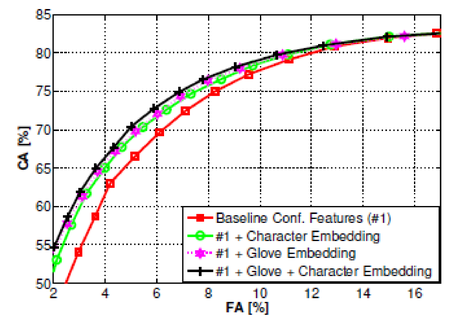
- Acoustic models use DNNs/LSTMs with 4-6 layers
- Consistent improvements across operating points from letters as wells as pronunciation embeddings
- The embeddings show similar gains for Server tasks, 8-10% FA relative reduction
- Phone embeddings are better for Xbox, 8% relative reduction in FA at CA=90%





Results in combination with Glove Embeddings

- Expanded our work to include NLP embeddings like Glove
- Proposed embeddings are complementary to Glove, and we see gain in combination with Glove



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.