Learning Joint Audio-Phonetic Spelling Embeddings from Noisy Labels (Speech Recognition)

Mohamed G. Mahmoud
elgeish@stanford.edu

1 Problem Description

We propose a building block for speech recognition tasks like automatic speech recognition (ASR) and audio search. The outcome will be a model that maps words to vectors in a shared latent space, which connects audio and phonetic modalities, such that words that sound similar — in either modality — end up clustered together in the space.

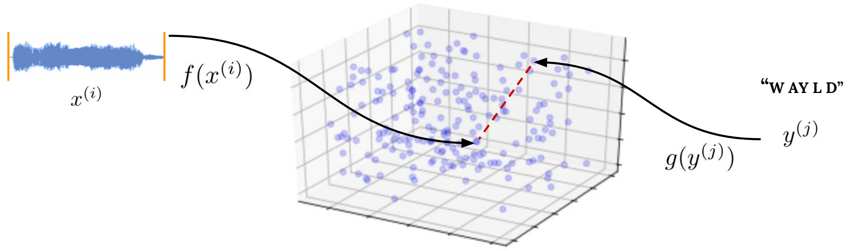

Figure 1: A naïve example of mapping input words, $x^{(i)}$ and $y^{(j)}$, to vectors in a shared latent space.

2 Challenges

The main challenge is mining hard examples for training. Training using words picked at random from a typical corpus may yield a model that almost always predicts that all word pairs are dissimilar (given the prior distribution). A more balanced class distribution is desired.

3 Dataset

The raw data consist of 30k 30-second, single-channel recordings and their respective transcripts. The recordings were captured at a sample rate of 16kHz and encoded using pulse-code modulation with 16-bit precision. Transcripts do not include word alignments and may include errors. The dataset is proprietary and growing. To obtain the phonetic spelling of words, we'll use The CMU Pronouncing Dictionary: It contains over 134k words and their pronunciations; out-of-vocabulary (OOV) words will excluded from training data (and used for qualitative evaluation).

4 Preprocessing

To align an audio segment with its respective transcript, we will attempt to force-align each transcript in the raw dataset with recognition hypotheses generated by an ensemble of ASR systems.

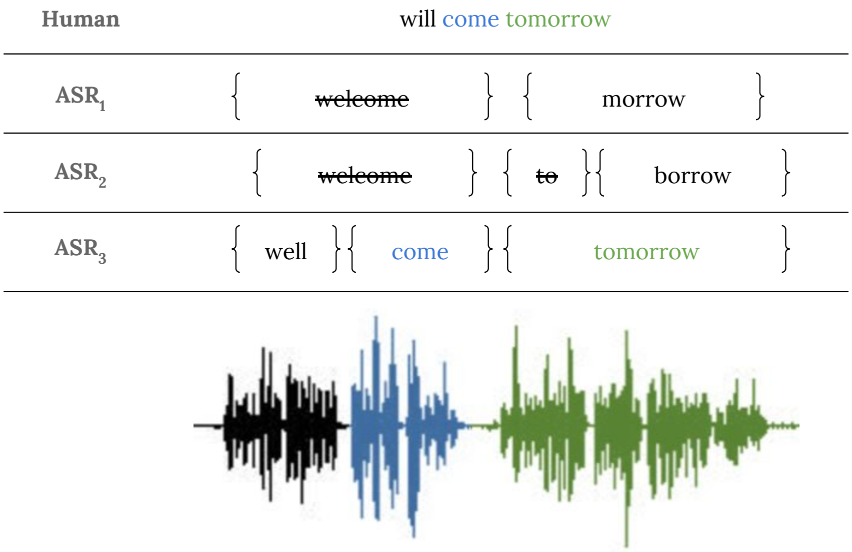

Figure 2: An example of a forced alignment that produced two same-word clusters (in blue and green) and discarded a few (stricken out) as too noisy. The words "well" and "borrow" will be used as hard negative examples for "will" and "tomorrow" respectively.

We aim at grouping pairs of words that sound the same as positive examples; ones that sound slightly different as hard negative examples; and ones that sound different as typical negative examples.

5 Learning Method

One promising method is weak supervision using noisy labels in order to learn how to bring similar-sounding pairs together while separating different-sounding ones. The objective is minimizing the loss between true and predicted cosine distances for each word pair given a margin (a function of the phonetic edit distance). We'll explore Siamese recurrent neural networks to train a model and learn the embeddings along the way. An existing implementation [6], which was trained using 100k same-word pairs, may be used as a baseline; we aim to improve previous results by training using far more data, thanks to noisy labels, and mining hard negative examples instead of randomly selected negative examples (distractors). We will also explore simpler implementations and related literature: [2], [1], [7], [3], [4], and [5].

6 Evaluation

In addition to our own test set, we will evaluate the embeddings using the same downstream classification task from [6] on the same data set (the Switchboard conversational English corpus): determining same- or different- word pairs and measuring performance via average precision (AP). By sweeping a threshold of the cosine distance between words, we obtain a precision-recall curve from which we compute the AP. We may include additional evaluations for the phonetic-spelling modality (e.g., comparing the phonetic edit distance of word pairs to the cosine distance of their respective embeddings). Quantitatively, we will perform error analysis and examine the AP of searching for OOV words.

References

[1] David Harwath and James R Glass. "Learning word-like units from joint audio-visual analysis". In: *arXiv preprint arXiv:1701.07481* (2017).

[2] Wanjia He, Weiran Wang, and Karen Livescu. "Multi-view recurrent neural acoustic word embeddings". In: *International Conference on Learning Representations* (2017).

[3] Herman Kamper, Weiran Wang, and Karen Livescu. "Deep convolutional acoustic word embeddings using word-pair side information". In: *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE. 2016, pp. 4950–4954.

[4] Keith Levin et al. "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings". In: *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE. 2013, pp. 410–415.

[5] Yingming Li, Ming Yang, and Zhongfei Mark Zhang. "A Survey of Multi-View Representation Learning". In: *IEEE Transactions on Knowledge and Data Engineering* (2018).

[6] S. Settle and K. Livescu. "Discriminative acoustic word embeddings: Tecurrent neural network-based approaches". In: *2016 IEEE Spoken Language Technology Workshop (SLT)*. Dec. 2016, pp. 503–510.

[7] Liwei Wang, Yin Li, and Svetlana Lazebnik. "Learning deep structure-preserving image-text embeddings". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 5005–5013.