SEMANTIC QUERY-BY-EXAMPLE SPEECH SEARCH USING VISUAL GROUNDING

Herman Kamper[1] *Aristotelis Anastassiou*[1] *Karen Livescu*[2]

[1]E&E Engineering, Stellenbosch University, South Africa & [2]TTI-Chicago, USA

kamperh@sun.ac.za, aristaki@me.com, klivescu@ttic.edu

ABSTRACT

A number of recent studies have started to investigate how speech systems can be trained on untranscribed speech by leveraging accompanying images at training time. Examples of tasks include keyword prediction and within- and across-mode retrieval. Here we consider how such models can be used for query-by-example (QbE) search, the task of retrieving utterances relevant to a given spoken query. We are particularly interested in *semantic* QbE, where the task is not only to retrieve utterances containing exact instances of the query, but also utterances whose meaning is relevant to the query. We follow a segmental QbE approach where variable-duration speech segments (queries, search utterances) are mapped to fixed-dimensional embedding vectors. We show that a QbE system using an embedding function trained on visually grounded speech data outperforms a purely acoustic QbE system in terms of both exact and semantic retrieval performance.

Index Terms— Multimodal modelling, visual grounding, semantic retrieval, query-by-example, speech search.

1. INTRODUCTION

While the field of speech processing has made great strides for tasks and domains with large amounts of available training data, lower-data domains and languages are still not adequately addressed. This has led many to explore alternative, weaker sources of supervision when labelled data is not available [1–3]. One form of weak supervision that has seen recent success is visual grounding: the use of images paired with speech data [4–8]. While we do not expect to be able to train a complete speech recognizer from unlabelled speech and images, it is possible to train models for more constrained tasks, such as cross-modal retrieval [5,6], unsupervised learning of word-like units [9, 10], keyword search [11], and semantic search [12].

Here we explore how unlabelled speech paired with visual context can be used for *semantic query-by-example search* (semantic QbE). Given a spoken query and a search database of spoken utterances, the task is to find utterances that are semantically relevant to the query. For example, given a spoken query like "children", we would like to retrieve utterances containing the word "children" but also utterances *about* children, like "two girls are playing hopscotch." This differs from standard QbE, which only seeks exact matches to the query; from keyword spotting and spoken term detection, where the query is written instead of spoken; and from typical semantic search tasks [12, 13], which also involve textual queries.

Our approach to semantic QbE is embedding-based: We learn an embedding function that maps from segments of speech—queries, search utterances, or sub-segments of search utterances—to fixed-dimensional vectors; we search for semantic matches by finding the minimum distance between query and search utterance embedding vectors. In this respect our approach is similar to those in recent embedding-based QbE work [14–17], and also some embedding-based spoken term detection work [18]. The key difference is that our embedding function must be learned in such a way that similar embedding vectors are *semantically* rather than *phonetically* similar. For this purpose, training on visually grounded speech data provides the source of semantic information.

Our setting and task are natural to consider for low-resource and even unwritten languages [19]. Like much prior work on low-resource methods, in this paper we use English-language data, but we do not use any transcriptions in order to simulate a low-resource language setting.

2. SEMANTIC QbE USING VISUAL GROUNDING

We perform (semantic) QbE using an embedding-based approach (§2.1), where the acoustic embedding function is obtained through visual grounding (§2.2). We consider two different ways to embed search utterances (§2.3 and §2.4).

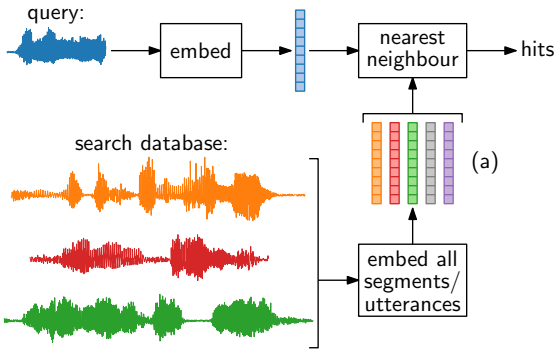

Fig. 1. Embedding-based query-by-example (QbE) search.

2.1. Embedding-based QbE

Traditional QbE [20–22] approaches are based on looking for alignments between the query and spans in the search database, most commonly using dynamic time warping (DTW). In contrast, embedding-based QbE [14–17] relies on an *acoustic embedding* function, which maps a speech segment (of variable length) to a fixed-dimensional vector. Ideally, instances of the same word should be mapped to similar vectors while unrelated words (or utterances) should have embeddings that are far apart. Instead of requiring an alignment between variable-duration segments (as in conventional QbE), queries and search utterances (or sub-segments of search utterances) are compared directly in this embedding space. The overall embedded QbE approach is illustrated in Figure 1.

Various acoustic embedding functions have been proposed, with neural models used in several studies [16, 18, 23–30]. Most of these methods, however, require labelled training data. For example, some recent work uses convolutional or recurrent neural networks learned by optimizing a contrastive loss using a set of known same-word pairs [26, 28]. Even for studies considering unsupervised acoustic embeddings, true word boundaries are normally used (e.g., in [24, 27]). Exceptions include [18, 31], which use no annotations.

Here we consider settings where no text labels or any other annotations (such as word boundaries) are available; instead, unlabelled speech is paired with visual context, which serves as the sole supervision signal. We use this visual information to train an acoustic embedding function for use in embedding-based QbE. This setting is relevant, e.g., for very low-resource languages or languages without an orthography [19].

2.2. A visually grounded model of speech as the acoustic embedding function

Given a corpus of images with spoken captions, neither having textual labels, our goal is to obtain a network that can map an arbitrary length speech segment to a fixed-dimensional vector. Many of the recently proposed vision+speech approaches can be used for this (§1). Here we use the method of [11, 12]. This approach (Figure 2) takes advantage of a separate visual tagger, which predicts relevant text labels for a given input image. The tagger produces soft keyword labels (posteriors of tags) for each training image in the audio-visual training set. These are then used as targets for a neural network that maps unlabelled speech to keyword labels. Without observing any transcriptions, the model can be used to predict which (written) words are present in a previously unseen input utterance, acting as a spoken bag-of-words classifier. This is not possible with most other vision+speech models, which map speech and images into a shared space but do not produce labels.

Formally, training image I is paired with spoken caption $X = \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$, where $\boldsymbol{x}_t$ is an acoustic feature vector, e.g. MFCCs, for frame t. An external vision system (Figure 2, left) is used to tag I with soft textual labels, giving targets

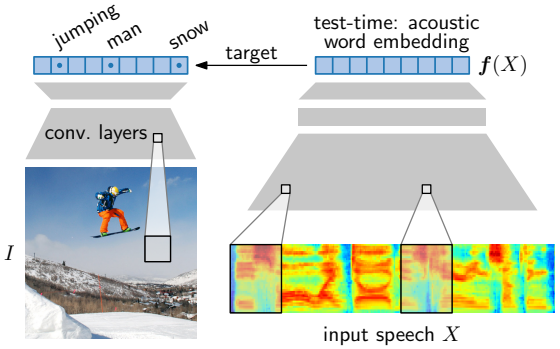

Fig. 2. The acoustic embedding CNN is trained using paired images and unlabelled spoken captions. Training targets for the speech network $\boldsymbol{f}(X)$ (right) is obtained from the external visual tagger (left).

$\hat{\boldsymbol{y}}_{\text{vis}} \in [0,1]^W$, with $\hat{y}_{\text{vis},w} = P(w|I)$ the estimated probability of word (tag) w being present in image I, and W the number of possible visual tags. Using $\hat{\boldsymbol{y}}_{\text{vis}}$ as target, we train the speech model $\boldsymbol{f}(X)$ (Figure 2, right). This model consists of a CNN over the speech X with a final sigmoidal layer so that $\boldsymbol{f}(X) \in [0,1]^W$. We interpret each dimension of the output as $f_w(X) = P(w|X)$, and train the model using the summed cross-entropy loss (see [12]). Note that $\boldsymbol{f}(X)$ is not a distribution over the output vocabulary since any number of keywords can be present in an utterance: it is a multi-label classifier where each dimension $f_w(X)$ can have any value in $[0,1]$. Also note that the size-W output vocabulary is implicitly specified by the visual tagger.

After training, $\boldsymbol{f}(\cdot)$ can be applied to unseen speech (without any visual input). For spoken input X of arbitrary duration, the network output $\boldsymbol{f}(X) \in [0,1]^W$ is a single W-dimensional vector, which we can use as the acoustic embedding for that input. We could also use representations from an intermediate layer in the network, which could be useful when a specific dimensionality is desired. We consider both options in §3. We can thus use $\boldsymbol{f}(\cdot)$ (Figure 2, right) directly as the embedding function for embedding-based QbE (Figure 1). A query (Figure 1, top left) can be fed to the $\boldsymbol{f}(\cdot)$ network and its embedding obtained. For embedding the search utterances (Figure 1, bottom left), we consider two options.

2.3. FAST: Embed and compare query and search utterances as single vectors

The first option is to feed an entire search utterance to $\boldsymbol{f}(\cdot)$, obtaining a single embedding for that utterance. To determine whether a query occurs in (or is relevant to) an utterance, the query embedding is compared to that single utterance embedding. Here we use cosine distance for this comparison. One disadvantage of this approach is that, even if an instance of the query occurs in the utterance exactly, the utterance embedding will also capture information from all the other words occurring in that utterance. We use normalization techniques to temper this effect (§3.3). The advantage of this approach is that it

is computationally very efficient, which is why we refer to it as FAST. For FAST, there is thus one embedding for every search utterance at (a) in Figure 1. The whole-utterance embedding approach has also been used, for example, in embedding-based (written) keyword search [18].

2.4. DENSE: Embed and compare queries to sub-segments within search utterances

Instead of obtaining a single embedding for an entire utterance, the DENSE method splits each utterance into overlapping segments from some minimum duration to some maximum duration. Each segment is then embedded separately using $\boldsymbol{f}(\cdot)$, as illustrated in Figure 3. This is similar to the approach used in some previous embedding-based QbE work [14, 16].[1]

To determine the relevance of a search utterance to a query, the query embedding is compared to all the embeddings from that utterance. Specifically, we compare the query to each of the utterance sub-segment embeddings using cosine distance, and then take the minimum cosine distance as the final score for the relevance of that query to the utterance. This approach is slower than FAST, but still much more efficient than performing full alignments between queries and search utterances using traditional DTW (see §3.2). DENSE can also predict the location of the segment within an utterance that resulted in a match. This is not directly possible with FAST, which scores entire utterances. However, we do not evaluate localization performance here, and leave this for future work. For DENSE, there will therefore be multiple embeddings for each search utterance at (a) in Figure 1, and this number will depend on the minimum and maximum segment duration and step size.

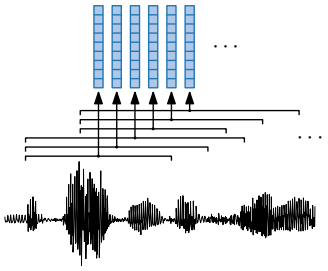

Fig. 3. For DENSE, search utterances are split into overlapping segments which are embedded individually.

3. EXPERIMENTS

3.1. Experimental setup and evaluation

We train our visually grounded acoustic embedding model (§2.2) on the corpus of parallel images and spoken captions of [32], containing 8000 images with 5 spoken captions each, divided into train, development and test sets. The audio comprises around 37 hours of active speech in total. 67 keyword types are selected randomly from transcriptions of the training portion of the corpus. In the development and test sets, spoken instances of these keywords are extracted as queries, while a disjoint part of each set is used as a search collection. There are multiple queries of the same type, with approximately 2000 spoken queries in total being matched to around 4000 search utterances (in each evaluation set). We parametrize speech as 13-dimensional MFCCs with first and second order derivatives. Utterances longer than 8 s are truncated to 8 s.

The structure and optimization procedure of the visually grounded embedding network (§2.2) are the same as in [12]: it uses three convolutational layers, one fully connected layer, and a final 1000-unit sigmoid output layer. We deal with the variable duration of utterances by pooling units over time at the last convolutional layer. Each of the $W = 1000$ units in the final output corresponds to an image tag from the external visual classifier (also see [12]). Note that all 67 of the keyword types occurs as one of the tags. The output of this speech network is used as embedding function in the FAST and DENSE matching variants. To explicitly denote that these systems use a visually grounded embedding method, we denote them as FASTGROUNDED and DENSEGROUNDED, respectively. For DENSE, a minimum segment duration of 200 ms and a maximum of 600 ms are used with a step of 30 ms.

As a baseline, we use a simple implementation of a DTW-based QbE system that performs successive alignments: a query is swept over a search utterance (30 ms step size), the DTW alignment cost is calculated over the overlapping segments (of the same length as the query), and the overall best alignment is taken as the score for how likely that utterance is to contain the query. More advanced DTW-based QbE systems have been proposed [22,33] (mainly to improve efficiency), but we restrict ourselves to this exhaustive-search implementation.

We use 3 metrics to quantify how well a QbE system predicts exact query matches [20, 21]: $P@10$ is the average precision (across keywords, in %) of the 10 highest-scoring proposals (utterances); $P@N$ is the average precision of the top N proposals, with N the number of true occurrences of the keyword; and equal error rate (EER) is the average error rate at which false acceptance and false rejection rates are equal.

For 1000 of the test utterances, semantic labels were collected in [12] using Amazon Mechanical Turk for the same set of 67 keyword types we use here. We use these labels to evaluate semantic QbE performance, where the goal is to retrieve all utterances that are semantically relevant, irrespective of whether an instance of the query occurs exactly in the utterance or not. Each of the 1000 test utterances was labelled by 5 annotators. By taking the majority decision, a hard label of whether an utterance is semantically relevant or irrelevant to a query can be assigned. We use these hard labels to calculate semantic $P@10$, $P@N$ and EER. We also calculate Spearman's ρ, which measures the correlation between a system's ranking and the actual number of annotators that marked a keyword as relevant to an utterance [34, 35].

[1] While [14, 16] use an approximate nearest neighbour search procedure, we use exhaustive search here.

Table 1. Exact and semantic QbE performance on test data. DTW performs full alignment. GROUNDED systems use acoustic embeddings trained only using visual supervision, while SUPERVISED systems are trained on text labels. FAST systems represent search utterances as single embeddings, while DENSE systems embed overlapping segments within search utterances.

		Exact QbE (%)			Semantic QbE (%)				Run-time (min)
	Model	$P@10$	$P@N$	EER	$P@10$	$P@N$	EER	Spearman's ρ	
Baselines:	RANDOM	4.5	4.5	50	9.5	9.1	50	5.9	-
	DTW	54.6	24.9	32.1	44.3	24.3	38.7	13.7	4080
Our systems:	FASTGROUNDED	27.5	17.9	38.9	32.6	23.2	41.4	12.8	< 1
	DENSEGROUNDED	56.0	37.3	21.7	55.5	37.3	30.0	14.9	621
Supervised:	FASTSUPERVISED	60.7	41.3	27.2	56.6	30.9	39.8	8.5	< 1
	DENSESUPERVISED	72.0	55.7	12.0	71.2	46.4	27.4	13.5	568

3.2. Results: Exact and semantic QbE

Table 1 shows exact and semantic QbE results. A random baseline is included for reference, which assigns a random relevance score for a search utterance. For the SUPERVISED systems, a speech network was trained on text transcriptions to perform keyword prediction, and embeddings taken from the final output. These systems therefore represent the case where perfect text labels are available for training utterances.

FASTGROUNDED is outperformed by the conventional DTW QbE approach across all metrics. The DENSEGROUNDED system, however, outperforms DTW across all metrics. This also comes with a speed benefit: DENSEGROUNDED is more than 6 times faster than DTW. (Run-time reported for embedding comparisons on a single CPU; we parallelized all systems.) FASTGROUNDED, which can compare a query and search utterance using a single comparison, is several orders faster than the other approaches, but comes with a cost in performance.

Comparing the exact QbE metrics to the semantic QbE metrics, we see that DTW and the SUPERVISED systems all perform worse on semantic QbE. In contrast, the GROUNDED systems perform better on all metrics when moving to semantic QbE. DENSEGROUNDED in fact achieves the best overall performance on Spearman's ρ, which takes the soft annotator scores into account. This also aligns with the findings in [12], which considered keyword spotting (where queries are written rather than spoken), and also found that the visually grounded systems aligned better with actual annotator counts.

3.3. Additional experiments

The DENSE systems were tuned on development data to set the maximum and minimum durations and step size of the segments (although, because of the run-time of these systems, extensive hyper-parameter optimization was not possible).

In order to determine what effect lower-dimensional embeddings would have, we also considered a visually grounded embedding network with a penultimate 256-dimensional bottleneck layer; using the bottleneck layer outputs as our acoustic embedding function $\boldsymbol{f}(\cdot)$ worsened $P@10$, $P@N$ and EER by between 5 and 10% absolute. Apart from cosine distance as a measure of embedding distance, we also considered Euclidean and Kullback-Leibler divergence, but cosine proved best.

Our original motivation for FASTGROUNDED was that, if a query contains a keyword of a particular type, the embedding from $\boldsymbol{f}(\cdot)$ will have a single dimension with a high probability (since in our case each embedding dimension corresponds to a particular visual tag and all query types occur as tags). By only considering this specific dimension for all of the search utterance embeddings, a quick retrieval would be possible. The reasonable performance of FASTSUPERVISED (Table 1) shows that this is in principle possible. But we found that when visual grounding is used, embeddings are highly influenced by the prior occurrence of specific visual tags. The embedding dimension corresponding to "man", for instance, typically has a high score (irrespective of the input), since many training images contain men. To alleviate this effect, we performed mean and variance normalization on all of the evaluation queries and search utterances using mean and variance estimates from the training embeddings. (We also considered several other normalization methods, but this approach proved most robust.)

4. CONCLUSION

For settings where annotated speech resources are not available, we have shown that query-by-example speech search (QbE) is possible using a model trained on images and unlabelled spoken captions. Such a model outperforms a conventional acoustic alignment-based (DTW) system, in terms of both exact QbE and semantic QbE, where the goal is to also retrieve non-verbatim matches related in meaning to the query. Here we used a specific vision+speech model, but we plan to also investigate how other models (e.g., [5]) can be used to obtain fixed-dimensional acoustic embeddings. There has also been recent work on acoustic-only methods for semantic-acoustic embedding [29, 36], which could prove complementary to our approach. Finally, we plan to consider how visual supervision can be used in truly low-resource languages.

We thank NVIDIA for sponsoring a Titan Xp GPU for this work. This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-18-1-0166.

5. REFERENCES

[1] G. Synnaeve, T. Schatz, and E. Dupoux, "Phonetics embedding learning with side information," in *Proc. SLT*, 2014.

[2] D. Palaz, G. Synnaeve, and R. Collobert, "Jointly learning to locate and classify words using convolutional networks," in *Proc. Interspeech*, 2016.

[3] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Low-resource speech-to-text translation," in *Proc. Interspeech*, 2018.

[4] G. Synnaeve, M. Versteegh, and E. Dupoux, "Learning words from images and speech," in *NIPS Workshop Learn. Semantics*, 2014.

[5] D. Harwath, A. Torralba, and J. R. Glass, "Unsupervised learning of spoken language with visual context," in *Proc. NIPS*, 2016.

[6] G. Chrupała, L. Gelderloos, and A. Alishahi, "Representations of language in a model of visually grounded speech signal," *Proc. ACL*, 2017.

[7] O. Scharenborg et al., "Linguistic unit discovery from multimodal inputs in unwritten languages: Summary of the "Speaking Rosetta" JSALT 2017 Workshop," *Proc. ICASSP*, 2018.

[8] H. Kamper and M. Roth, "Visually grounded cross-lingual keyword spotting in speech," *Proc. SLTU*, 2018.

[9] D. Harwath and J. R. Glass, "Learning word-like units from joint audio-visual analysis," in *Proc. ACL*, 2017.

[10] D. Harwath et al., "Jointly discovering visual objects and spoken words from raw sensory input," *arXiv preprint arXiv:1804.01452*, 2018.

[11] H. Kamper, S. Settle, G. Shakhnarovich, and K. Livescu, "Visually grounded learning of keyword prediction from untranscribed speech," in *Proc. Interspeech*, 2017.

[12] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic speech retrieval with a visually grounded model of untranscribed speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 27, no. 1, pp. 89–98, 2019.

[13] L.-s. Lee, J. Glass, H.-y. Lee, and C.-a. Chan, "Spoken content retrieval—beyond cascading speech recognition with text retrieval," *IEEE Trans. Audio, Speech, Language Process.*, vol. 23, no. 9, pp. 1389–1420, 2015.

[14] K. Levin, A. Jansen, and B. Van Durme, "Segmental acoustic indexing for zero resource keyword search," in *Proc. ICASSP*, 2015.

[15] G. Chen, C. Parada, and T. N. Sainath, "Query-by-example keyword spotting using long short-term memory networks," in *Proc. ICASSP*, 2015.

[16] S. Settle, K. Levin, H. Kamper, and K. Livescu, "Query-by-example search with discriminative neural acoustic word embeddings," in *Proc. Interspeech*, 2017.

[17] Y.-H. Wang, H.-y. Lee, and L.-s. Lee, "Segmental audio word2vec: Representing utterances as sequences of vectors with applications in spoken term detection," in *Proc. ICASSP*, 2018.

[18] K. Audhkhasi, A. Rosenberg, A. Sethy, B. Ramabhadran, and B. Kingsbury, "End-to-end asr-free keyword search from speech," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1351–1359, 2017.

[19] G. Adda et al., "Breaking the unwritten language barrier: The BULB project," *Proc. SLTU*, 2016.

[20] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Proc. ASRU*, 2009.

[21] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *Proc. ASRU*, 2009.

[22] A. Jansen and B. Van Durme, "Indexing raw acoustic features for scalable zero resource search," in *Proc. Interspeech*, 2012.

[23] A. L. Maas, S. D. Miller, T. M. O'Neil, A. Y. Ng, and P. Nguyen, "Word-level acoustic modeling with convolutional vector regression," in *Proc. ICML Workshop Representation Learn.*, 2012.

[24] K. Levin, K. Henry, A. Jansen, and K. Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings," in *Proc. ASRU*, 2013.

[25] S. Bengio and G. Heigold, "Word embeddings for speech recognition," in *Proc. Interspeech*, 2014.

[26] H. Kamper, W. Wang, and K. Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," in *Proc. ICASSP*, 2016.

[27] Y.-A. Chung, C.-C. Wu, C.-H. Shen, and H.-Y. Lee, "Unsupervised learning of audio segment representations using sequence-to-sequence recurrent neural networks," *in Proc. Interspeech*, 2016.

[28] S. Settle and K. Livescu, "Discriminative acoustic word embeddings: Recurrent neural network-based approaches," in *Proc. SLT*, 2016.

[29] Y.-A. Chung and J. R. Glass, "Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech," in *Proc. Interspeech*, 2018.

[30] N. Holzenberger, M. Du, J. Karadayi, R. Riad, and E. Dupoux, "Learning word embeddings: Unsupervised methods for fixed-size representations of variable-length speech segments," *in Proc. Interspeech*, 2018.

[31] Y.-A. Chung, W.-H. Weng, S. Tong, and J. R. Glass, "Unsupervised cross-modal alignment of speech and text embedding spaces," in *Proc. NIPS*, 2018.

[32] D. Harwath and J. R. Glass, "Deep multimodal semantic embeddings for speech and images," in *Proc. ASRU*, 2015.

[33] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 186–197, 2008.

[34] E. Agirre et al., "A study on similarity and relatedness using distributional and wordnet-based approaches," in *Proc. HLT-NAACL*, 2009.

[35] F. Hill, R. Reichart, and A. Korhonen, "SimLex-999: Evaluating semantic models with (genuine) similarity estimation," *Comput. Linguist.*, vol. 41, no. 4, 2015.

[36] Y.-C. Chen, S.-F. Huang, C.-H. Shen, H.-y. Lee, and L.-s. Lee, "Phonetic-and-semantic embedding of spoken words with applications in spoken content retrieval," *arXiv preprint arXiv:1807.08089*, 2018.