Rapid Evaluation of Speech Representations for Spoken Term Discovery

Michael A. Carlin, Samuel Thomas, Aren Jansen, Hynek Hermansky

Center for Language and Speech Processing
Human Language Technology Center of Excellence
Department of Electrical and Computer Engineering
Johns Hopkins University, Baltimore, MD 21218, USA
{macarlin,samuel,aren,hynek}@jhu.edu

Abstract

Acoustic front-ends are typically developed for supervised learning tasks and are thus optimized to minimize word error rate, phone error rate, etc. However, in recent efforts to develop zero-resource speech technologies, the goal is not to use transcribed speech to train systems but instead to discover the acoustic structure of the spoken language automatically. For this new setting, we require a framework for evaluating the quality of speech representations without coupling to a particular recognition architecture. Motivated by the spoken term discovery task, we present a dynamic time warping-based framework for quantifying how well a representation can associate words of the same type spoken by different speakers. We benchmark the quality of a wide range of speech representations using multiple frame-level distance metrics and demonstrate that our performance metrics can also accurately predict phone recognition accuracies.

Index Terms: evaluation methods, acoustic front-end, spoken term discovery, zero resource

1. Introduction

In the zero resource setting, where we consider tasks such as keyword spotting, topic identification, and spoken term discovery, we are without the requisite transcribed speech to train recognizers. Here, the usual evaluation metrics of phone and word error rate are of little use since we are unable to construct the pipeline needed to measure them. At the recognition back-end, language modeling research has long employed perplexity as an easily computable stand-in for word error rate in the absence of a full-blown recognizer; however, the acoustic front-end has no such recognizer-independent metric in wide use. Thus, our goal is to define a more basic means to evaluate the potential of an acoustic front-end for enabling linguistically motivated pattern recognition downstream. The fundamental role of an acoustic front-end is to map speech signals containing the same word or phone spoken by multiple speakers to similar trajectories (as a vector time series) in some acoustic feature space. Hence, without committing to a subsequent acoustic modeling technique, we require a means to directly quantify separability of same-word and different-word trajectory pairs across a wide range of speakers and word types.

As our motivating application, we consider the task of spoken term discovery, whose goal is to automatically discover repeated words and phrases in continuous speech. Recently proposed approaches to the problem [1, 2, 3, 4, 5] have relied on variants of segmental dynamic time warping (S-DTW), which attempts to efficiently perform a dynamic time warping search of a speech collection against itself to discover repeated trajectories up to some reasonable amount of nonlinear internal time warping. If we factor out the segmental aspect of the algorithm by providing presegmented word examples, we can test the quality of a speech representation for the task by using normal dynamic time warping comparisons.

In particular, given a modest collection of presegmented word examples drawn from a multispeaker speech corpus, we (i) compute the dynamic time warping (DTW) distance between all example pairs, and (ii) quantify how well the DTW distances predict same or different word type using standard information retrieval metrics. By this measure, improved performance implies improved segmental DTW performance, enabling more efficient discovery of repeated terms in speech. Moreover, this evaluation completes in minutes using a typical university compute cluster, enabling rapid feedback if optimizing front-end architecture parameters (e.g. number of filters, linear prediction model order, target reduced dimension, etc.).

While motivated directly by the previous term discovery approaches that rely on DTW, the proposed metric appeals to fundamental pattern recognition principles underlying speaker independent recognition of phones and words. In particular, after optimal alignment, front-end representations that reduce distance between corresponding frames across the word examples of the same type and simultaneously increase distance for different types can only benefit downstream supervised tasks. Thus, we expect the proposed performance metrics to provide a useful tool to rapidly evaluate front-end quality, not only for the spoken term discovery task, but for supervised ASR applications more generally. As an example, we demonstrate that when applied to phonetic posteriorgrams of varying quality, the DTW-based performance metrics are nearly perfectly correlated with downstream phone recognition. We begin with a detailed description of the evaluation framework.

2. Evaluation Description

We assume we are provided with a set of M term candidates $\mathcal{W} = \{w_i\}_{i=1}^{M}$. Let each word segment $w_i = x_1 x_2 \dots x_{T_i}$, $x_t \in \mathbb{R}^d$ be a sequence of observed acoustic vectors. We consider the following four sets of word pairs $(w_i, w_j) \in \mathcal{W} \times \mathcal{W}, i \neq j$:

- $\mathcal{C}_1$: Same word, same speaker (SWSP)
- $\mathcal{C}_2$: Same word, different speakers (SWDP)
- $\mathcal{C}_3$: Different word, same speaker (DWSP)
- $\mathcal{C}_4$: Different word, different speakers (DWDP)

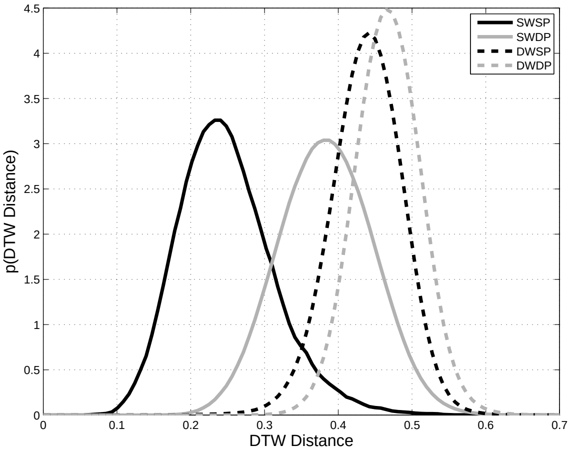

Figure 1: *DTW distance distributions for the four word pair categories $\{\mathcal{C}_k\}$ for normalized MFCCs.*

We expect that $|\mathcal{C}_1| \ll |\mathcal{C}_2| \ll |\mathcal{C}_3| \ll |\mathcal{C}_4|$ since we will have far fewer instances of the same word spoken by the same speaker as opposed to instances of different words spoken by different speakers.

We measure similarity between word pairs using the minimum DTW alignment cost $DTW(w_i, w_j)$ between w_i and w_j. DTW was originally applied in early speech recognizers to normalize for temporal variations between an observed vector time series and a stored template for tasks such as isolated word recognition. As mentioned earlier, a number of recent studies have successfully adapted the DTW algorithm to discover repeated segments in a speech utterance. Therefore, we expect that improvements realized across different sets of acoustic features in the context of this evaluation will necessarily benefit DTW-based speech pattern discovery.

Computing DTW requires we specify a distance metric between constituent frames in the word examples. For this evaluation we consider (1) Euclidean distance, (2) cosine distance, and (3) symmetric KL-Divergence. Euclidean and cosine distance are appropriate for directly comparing frames of raw acoustic data, whereas the symmetric KL-Divergence is meaningful when assessing similarity between probability distributions such as those produced by multi-layer perceptron-based phonetic acoustic models.

Figure 1 shows the kernel density estimates of the distribution of DTW distances using the cosine metric for each of the four sets of word pairs $\{\mathcal{C}_k\}$ as described above. The results are shown for standard 39-dimensional MFCCs from English words extracted from the Switchboard corpus. As expected, under $DTW(\cdot,\cdot)$ instances of the same word uttered by the same speaker (SWSP) are more similar to one another than instances of different words uttered by the same speaker (DWSP). We also note that the SWDP distribution falls between the SWSP and two DW distributions; the closer it is to SWSP the more speaker independent the representation. The ideal acoustic representation for ASR tasks is one for which the SWSP and SWDP distributions coincide and are well-separated from both the DWSP and DWDP curves.[1]

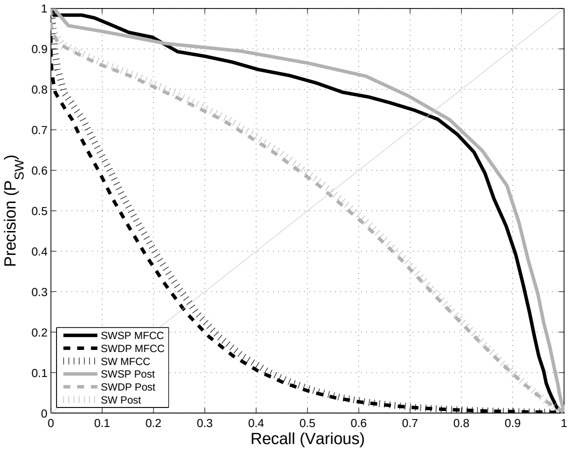

Figure 2: *Precision vs. recall for normalized MFCCs and 200-hour English posteriorgrams.*

We would like to simultaneously assess the degree of speaker independence and word discriminability of a representation using the information contained in the distributions shown in Figure 1. One possibility would be statistical measures of distributional separation, but early experiments indicated that the skewness in the distributional forms can lead to instability across feature sets. However, given that our target task is spoken term discovery, it is more appropriate to consider an evaluation in the context of information retrieval (IR), where the goal is to retrieve same word pairs from different word pair impostors. In particular, for a word pair $(w_i, w_j), i \neq j$, and some threshold τ, we declare that w_i and w_j correspond to the same word if $DTW(w_i, w_j) \leq \tau$. As we sweep the threshold τ we can sample a standard precision-recall curve, which we characterize using two criteria:

1. *Precision-Recall Breakeven* (PRB): The precision/recall at the operating point where the two quantities are equal.
2. *Average Precision* (AP): The area under the precision-recall curve, characterizing the average performance of the IR system across all operating points.

We can further split the same word IR task into subcases for SWSP and SWDP retrieval, allowing us to characterize the level of speaker dependence of a representation. To proceed, we first define

$$N_k(\tau) := |\{(w_i, w_j) \in \mathcal{C}_k : DTW(w_i, w_j) \leq \tau\}|$$

as the number of word pairs in $\mathcal{C}_k$ for which the DTW distance between segments is small enough so as to be declared as the same word. We can then define various precision and recall quantities involved in PRB and AP as follows:

$$P_{SW}(\tau) = \frac{N_1(\tau)}{\sum_{k=1}^{4} N_k(\tau)}, \qquad R_{SW}(\tau) = \frac{N_1(\tau) + N_2(\tau)}{|\mathcal{C}_1| + |\mathcal{C}_2|}$$
$$R_{SWSP}(\tau) = \frac{N_1(\tau)}{|\mathcal{C}_1|}, \qquad R_{SWDP}(\tau) = \frac{N_2(\tau)}{|\mathcal{C}_2|}$$

Here, P_{SW} and R_{SW} are the precision and recall of retrieving same words regardless of speaker; R_{SWSP} and R_{SWDP} are the recall for same and different speaker subcases, respectively.

[1] These criteria can be easily adapted for other tasks such as speaker identification, where maximal separation between SWSP and SWDP would be desirable.

Table 1: *Evaluation corpus word statistics.*

Word Count	Word Pairs	SWSP	SWDP	DWSP	DWDP
Switchboard English					
11.0k	60.7M	3k	93k	383k	60.3M
Fisher Spanish					
10.9k	56.7M	9k	90k	1.8M	54.8M

Figure 2 shows the precision and recall curves for standard 39-dimensional MFCCs and English MLP-based posteriorgrams (see Section 3), using a set of English words from the Switchboard corpus. We show P_{SW} versus the three recall scenarios just defined. A good speaker-dependent representation is indicated by a large breakeven PRB_{SP} of P_{SW} and R_{SWSP}, which is exhibited by both MFCCs and posteriorgrams. Likewise, a good speaker-independent representation produces a large breakeven PRB_{DP} between P_{SW} and R_{SWDP}; here, highly supervised posteriorgrams naturally outperform the raw acoustic features. Since the SWDP pairs vastly outnumber the SWSP pairs, the combined precision recall curve (P_{SW} vs. R_{SW}) is very similar to the same word, different speaker case. We compute average precision (AP) using the P_{SW} and R_{SW} (i.e., we do not consider the same/different speaker distinction). Note that since R_{SW} is dominated by SWDP pairs, AP largely characterizes the speaker independence of the representation.

3. Experiments

3.1. Evaluation Setup

The features we consider here can be broadly divided into two categories: (1) features that represent the underlying spectral properties of the speech signal and (2) high-resource, data-driven features that are derived from large amounts of data. Traditional acoustic features such as MFCCs and PLPs are derived by analyzing the short-time spectrum of speech (∼10–30 ms). More recently, short-term features have been derived by considering sub-band temporal envelopes estimated using frequency domain linear prediction (FDLP) [6]. As with the traditional features, temporal context is accounted for by augmentation with derivatives of the spectral trajectory at each frame.

Next, we consider high-resource, data-driven features derived by discriminatively training a multi-layer perceptron (MLP) on large amounts of acoustic features to estimate phone class posterior probabilities [7]. By considering multiple frames of acoustic features when training each MLP, posterior features also capture information in a larger temporal context. Furthermore, it has been observed that these features exhibit less speaker variability and perform better than the conventional features for various ASR tasks [8]. We train separate MLPs on English and Spanish with increasing amounts of conversational telephone speech (CTS). The English CTS database consists of more than 200 hours of conversational speech from the LDC English Switchboard and CallHome corpora. The phone labels are obtained by force aligning the word transcripts to the previously trained HMM/GMM models and we use a set of 45 phones [9]. The Spanish CTS database is made up of 200 hours of conversational speech from the LDC Spanish Switchboard and CallHome corpora. Phone labels for this database are obtained by force aligning word transcripts using BBN's Byblos recognition system and is labeled using 27 phones.

To estimate phone posteriors, 39-dimensional PLP features with a 9-frame context are used to train three-layer MLPs. The network is trained using the standard back propagation algorithm with cross entropy error criteria. The number of nonlinear hidden units increases with the amount of training data to learn more structure from the data. Table 3 describes the different network configurations for both languages along with their corresponding phone recognition accuracies on a 1.5 hour held-out data set; network complexity is indicated here by (input feature dimension × hidden layer size × number of phone classes).

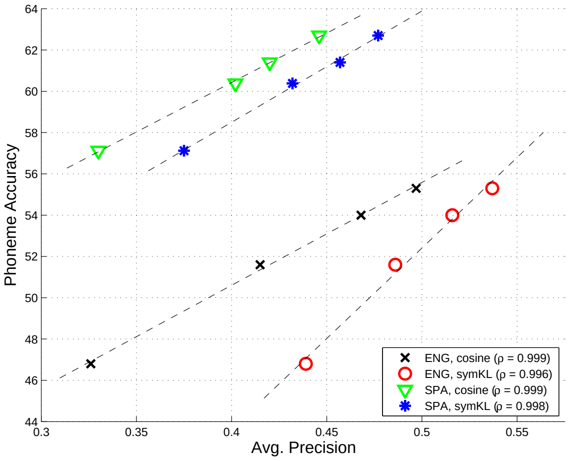

Figure 3: *Average precision vs. phone recognition accuracy.*

As a second data-driven feature type, we also consider a frame-level tokenization based on two vector quantization algorithms applied to raw acoustic features: traditional k-Means and self-learning vector quantization [10] (a technique developed for the zero resource setting). By labeling each acoustic observation with respect to a VQ codebook index (or codeword), we expect that computing DTW distance between integer-valued word segments will significantly reduce computation time. In this scenario, we consider binary-valued distances at the frame level of the form $d(x_i, x_j) = \mathbf{1}_{[x_i == x_j]}$ when computing $DTW(w_i, w_j)$.

For the acoustic and posteriorgram features, we consider our evaluation for both English speech from the Switchboard corpus and Spanish speech from the Fisher corpus. We define our word set $\mathcal{W}$ by drawing a random subset of conversations and extracting word examples at least 0.5 sec in length and five characters as text. A summary of the word pair statistics for each of the sets $\mathcal{C}_k$ described in Section 2 is given in Table 1.

3.2. Results and Discussion

Results of the evaluation on the raw acoustic features are shown in Table 2. We consider baseline and normalized MFCCs, PLPs, and spectral FDLP (FDLP-S) features; normalized here refers to shifting and scaling the features so that they are zero mean and unit variance. We also consider both Euclidean and cosine distance metrics, with cosine emerging as the clear winner. We find that despite the differences in ranges of scores across languages, normalization of the feature dimensions to zero mean/unit variance results in higher scores, which is consistent with conventional wisdom in the supervised setting. Interestingly, the rankings across the six features are preserved regardless of language.

Table 3 lists the evaluation results as well as phone recognition accuracies for posterior features learned from increasing amounts of training idea. We consider DTW distances using both cosine distance and symmetric KL-divergence. A few general trends are clear. First, we observe that all evaluation measures increase with the amount of training data available to the MLP, with each of the nets benefiting from an abundance

Table 2: *Raw acoustic feature results.*

Feature	Cosine (English)			Cosine (Spanish)			Euclidean (English)		
	PRB_{SP}	PRB_{DP}	AP	PRB_{SP}	PRB_{DP}	AP	PRB_{SP}	PRB_{DP}	AP
MFCC	0.629	0.226	0.163	0.326	0.117	0.064	0.592	0.212	0.131
MFCC (Norm)	0.738	0.255	0.191	0.465	0.137	0.091	0.659	0.228	0.145
PLP	0.474	0.180	0.112	0.266	0.122	0.069	0.512	0.198	0.123
PLP (Norm)	0.712	0.243	0.177	0.423	0.121	0.079	0.651	0.216	0.137
FDLP-S	0.403	0.166	0.099	0.209	0.101	0.049	0.455	0.186	0.114
FDLP-S (Norm)	0.714	0.281	0.215	0.454	0.154	0.106	0.669	0.247	0.169

Table 3: *Matched language posteriorgram results.*

Posteriorgram	Net complexity	Cosine			Symmetrized KL			Rec. Acc.
		PRB_{SP}	PRB_{DP}	AP	PRB_{SP}	PRB_{DP}	AP	
English 10 hr	351x1000x45	0.597	0.390	0.326	0.728	0.471	0.439	46.8
English 50 hr	351x2500x45	0.648	0.455	0.415	0.731	0.505	0.486	51.6
English 100 hr	351x5000x45	0.666	0.492	0.468	0.742	0.527	0.516	54.0
English 200 hr	351x7000x45	0.682	0.514	0.497	0.751	0.542	0.537	55.3
Spanish 10 hrs	351x1000x27	0.503	0.371	0.330	0.585	0.403	0.375	57.1
Spanish 50 hrs	351x2500x27	0.559	0.434	0.402	0.622	0.450	0.432	60.4
Spanish 100 hrs	351x3500x27	0.574	0.448	0.420	0.636	0.467	0.457	61.4
Spanish 200 hrs	351x4500x27	0.588	0.468	0.446	0.647	0.485	0.477	62.7

Table 4: *Unmatched posteriorgram results (using KL).*

Posteriorgram	Test Lang	PRB_{SP}	PRB_{DP}	AP
English 10 hrs	Spanish	0.422	0.210	0.155
English 50 hrs		0.423	0.217	0.162
English 100 hrs		0.414	0.219	0.162
English 200 hrs		0.411	0.215	0.158
Spanish 10 hrs	English	0.450	0.172	0.103
Spanish 50 hrs		0.365	0.157	0.085
Spanish 100 hrs		0.321	0.129	0.067
Spanish 200 hrs		0.317	0.133	0.073

of training data. For both languages in all training conditions, symmetric KL-divergence yields the best scores, a fact that has not been yet exploited in posterior-based spoken term discovery systems. More importantly, as observed in the table and illustrated in Figure 3, we see that the evaluation scores are nearly perfectly correlated with phone recognition accuracies, indicating potential for the proposed evaluation to predicting downstream ASR performance.

In previous spoken term discovery research [5], mismatched language posteriorgrams were used for the zero resource setting. Table 4 lists the cross-language performance of the posterior features, where English MLPs were used to represent Spanish speech and vice versa. The evaluation scores reflect a high degree of language specialization of the MLPs and increasing the amount of training data does not necessarily improve the effectiveness of the representation (and may in fact reduce evaluation performance).

Finally, we considered tokenizing MFCCs (normalized) with k-Means and SLVQ with results shown in Table 5. The reported results seem to reflect how variations in the codebook size (in terms of number of centroids/codewords) control the degree of speaker independence afforded by the representation (indicated by PRB_{DP}). Note that while the VQ results fall short of the other feature types discussed above, tokenized representations are extremely useful for downstream algorithmic gains. Thus, our proposed evaluation provides a new means for optimizing VQ algorithm parameters.

4. Conclusions

We have presented a framework for evaluating acoustic features in the context of a spoken term discovery task. In our study, we have demonstrated (i) the superiority of cosine distance and normalization for raw acoustic features, (ii) the superiority of symmetric KL divergence for posterior features, and (iii) in mismatched language conditions, increasing training data does not necessarily help. Moreover, we demonstrated for high-resource posteriorgrams, average precision is almost perfectly correlated with phone recognition accuracy, indicating our proposed method's potential for predicting performance of downstream recognition tasks.

Table 5: *Vector quantization results.*

VQ Type	Centroids	PRB_{SP}	PRB_{DP}	AP
k-Means	300	0.280	0.103	0.046
	1024	0.322	0.098	0.042
SLVQ	125	0.292	0.104	0.047
	327	0.304	0.101	0.045
	604	0.321	0.099	0.044
	721	0.237	0.078	0.029

5. References

[1] A. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE T-ASLP*, vol. 16, no. 1, pp. 186–197, 2008.

[2] L. ten Bosch and B. Cranen, "A computational model for unsupervised word discovery," in *Interspeech*, 2007.

[3] A. Muscariello, G. Gravier, and F. Bimbot, "Audio keyword extraction by unsupervised word discovery," in *Interspeech*, 2009.

[4] Y. Zhang and J. R. Glass, "Towards multi-speaker unsupervised speech pattern discovery," in *Proc. of ICASSP*, 2010.

[5] A. Jansen, K. Church, and H. Hermansky, "Towards spoken term discovery at scale with zero resources," in *Interspeech*, 2010.

[6] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," *IEEE Signal Processing Letters*, pp. 681–684, 2008.

[7] M. Richard and R. Lippmann, "Neural network classifiers estimate bayesian a posteriori probabilities," *Neural Computation*, vol. 3, pp. 461.483, 1991.

[8] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke, "On using MLP features in LVCSR," in *Proc. of INTERSPEECH*, 2010.

[9] T. Hain et al., "The AMI system for the transcription of speech in meetings," in *Proc. of ICASSP*, 2007.

[10] O. Rasanen, U. Laine, and T. Altosaar, "Self-learning vector quantization for pattern discovery from speech," in *Proc. of INTERSPEECH*, 2010.