Towards Unsupervised Training of Speaker Independent Acoustic Models

Aren Jansen,[1,2] *Kenneth Church*[1,3]

[1]Human Language Technology Center of Excellence,
[2]Department of Electrical and Computer Engineering, [3]Department of Computer Science
Johns Hopkins University, Baltimore, Maryland
aren@jhu.edu, kenneth.church@jhu.edu

Abstract

Can we automatically discover speaker independent phoneme-like subword units with zero resources in a surprise language? There have been a number of recent efforts to automatically discover repeated spoken terms without a recognizer. This paper investigates the feasibility of using these results as constraints for unsupervised acoustic model training. We start with a relatively small set of word types, as well as their locations in the speech. The training process assumes that repetitions of the same (unknown) word share the same (unknown) sequence of subword units. For each word type, we train a whole-word hidden Markov model with Gaussian mixture observation densities and collapse correlated states across the word types using spectral clustering. We find that the resulting state clusters align reasonably well along phonetic lines. In evaluating cross-speaker word similarity, the proposed techniques outperform both raw acoustic features and language-mismatched acoustic models.

Index Terms: speaker independent acoustic models, unsupervised training, spectral clustering

1. Introduction

Children are rarely presented in their development with isolated phonemes in supervised form. Instead they are exposed to words (some labeled by context, some not) produced by their caregivers and their explicit knowledge of phonetics is implicitly derived over time from segmental contrasts between the word types. Standard acoustic model training circumvents this process by using a pronunciation dictionary to map each word to a canonical phonetic pronunciation. When provided with orthographic word transcripts for a collection of training data, a speaker-independent phonetic acoustic model can be learned by using expectation-maximization to align the speech frames with the underlying phonetic sequence provided by the dictionary. However, in an extreme case of resource impoverishment, where you have a collection of untranscribed audio, but no transcripts or dictionaries, how might you discover this speaker-independent phonetic structure automatically?

There have been several recent bursts of activity surrounding unsupervised training of subword unit acoustic models for speech technologies. In [1, 2, 3, 4, 5], various strategies are proposed for unsupervised training of subword unit hidden Markov models (HMM) and Gaussian mixture models using a range of architectures. These techniques have been applied to a range of practical problems, including low resource keyword spotting, spoken term discovery, topic identification, automatic pronunciation lexicon generation (when provided transcripts only), and phonetic recognition. However, none of these approaches impose any explicit constraints in the training procedure to ensure the subword unit models they learn will provide consistent decoding of similar phonetic content *across speaker*. With this consideration in mind, this paper investigates whether achieving speaker independence requires the explicit definition of the phonetic structure via a transcript and dictionary, or whether weaker (unlabeled) word-level matching constraints, combined with appropriate subword unit clustering methods, can enable unsupervised training of a phone-like acoustic model that provides consistent output across speaker.

Similarly motivated past research investigated a related joint optimization of the unit inventory, pronunciation dictionary, and acoustic models in the service of improved *supervised* recognition training [6]. However, we propose a *fully unsupervised* process that relies on recent research efforts [7, 8, 9, 5, 10] in spoken term discovery that exploit the salience of long, word- and phrase-level patterns to search collections of untranscribed speech for repeated terms. These term discovery algorithms can produce a collection of word clusters, each consisting of several acoustic realizations of a given word type. Moreover, provided sufficient amount of data to search, each cluster can easily span speaker and gender, either through direct or transitive acoustic matches. While we do not know the phonetic structure underlying each cluster, we know that it should be similar for each example it contains. We exploit this weak constraint by (i) training for each word cluster a whole-word HMM with Gaussian mixture emission densities, and (ii) clustering the resulting subword states across word models to arrive at a phone-like speaker independent acoustic model. Using a recognizer independent evaluation metric, we find that even when starting with clusters containing only a small fraction of the examples present, our unsupervised acoustic model permits better cross-speaker word matching than alternative zero resource methods.

2. Unsupervised Training Procedure

At a high level, our vision for unsupervised acoustic model training rests on a two stage process of discovering long repeated patterns in speech (on the order of a word or short phrase) and using those repeated patterns to provide speaker independence constraints on subword acoustic model training. The architecture schematic is shown in Fig. 1 and consists of the follow steps: (1) Given raw acoustic features for a collection of speech (e.g. PLP or MFCCs), run a spoken term discovery procedure of the form presented in [7], producing word example clusters for a collection of n unspecified word types. The quantity and size of these clusters will scale inversely with the desired purity. (2) Train a whole-word model for each cluster in terms of a sequence of subword unit states. The collection of these subword states can be used to define a high-dimensional posteriorgram over a set of non-mutually exclusive

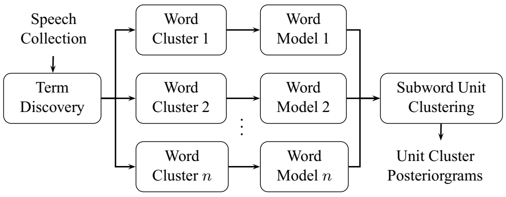

Figure 1: *Training procedure schematic.*

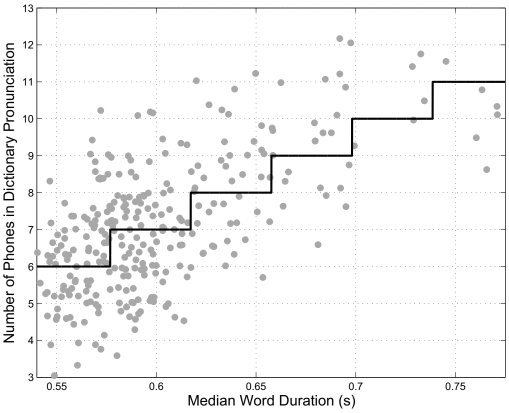

Figure 2: *Scatterplot of pronunciation phones versus the median word duration (values jittered for interpretability).*

units. (3) Cluster the collection of word-specific subword unit states across the word models and collapse posteriors accordingly, producing improved context-independent posteriorgrams. Here, speaker independence of the word models imply speaker independence of the subword unit posteriorgram representation. (4, *Optional*) Use the resulting posteriorgrams in place of the raw acoustic features and repeat as desired (not studied here).

For the purposes of this paper, we assume the term discovery task of Step 1 has already been performed with varying degrees of efficiency and thus take as given the word clusters for subsequent subword unit training; the challenges remaining in developing spoken term discovery systems will be addressed in Sec. 3.3. The goal of this paper, then, is limited to a demonstration that, in conjunction with a spoken term discovery algorithm, steps 2 and 3 are a zero resource stand-in for the traditional supervised acoustic model training strategy that relies on orthographic word transcripts and pronunciation dictionaries.

2.1. Learning Word-Specific Acoustic Models

Applied to a large collection of speech, we assume a spoken term discovery procedure has identified collection of word clusters for some vocabulary $\mathcal{W}$, where $|\mathcal{W}| = n$. The cluster for each $w \in \mathcal{W}$ contains N_w examples of the form $X_i^w = x_1 x_2 \ldots x_{T_i}$ for $i = 1, \ldots, N_w$, where T_i is the number of frames in the i-th example and each $x_t \in \mathbb{R}^d$ is the d-dimensional acoustic feature vector representing the t-th frame. We assume each word $w \in \mathcal{W}$ consists of some number Q_w of subword units and thus model each word with Q_w-state left-to-right HMM with 8-component Gaussian mixture model (GMM) observation densities (diagonal covariance) for each state.

Since we make no claim to the lexical identity of each word cluster, we have no immediate basis for the selection of an appropriate number of HMM states. However, if our goal is to recover phone-like states, we can attempt to choose Q_w according to median duration of examples in the cluster (i.e. the median value of T_i normalized by the frame rate) and the expected duration of a phone. Fig. 2 plots the number of phones present in the dictionary pronunciation versus the median word duration for 300 words. We find that there is a clear correlation and we can exploit this fact to predict a reasonable number of HMM states for each word cluster we are presented with. To accomplish this, we simply perform a linear regression and round the result to the nearest integer producing the step function overlaid in Fig. 2. While this method can under- or overpredict by as much as three phones, we note that in conversational speech the canonical pronunciation can be notoriously inaccurate.

Once settled on a model architecture, we can proceed by employing standard expectation-maximization training to estimate both the GMM parameters (component priors, means, and covariance matrices) and HMM parameters (state transition probabilities, which we discard) for each word $w \in \mathcal{W}$ using the training examples $\{X_i^w\}_{i=1}^{N_w}$. The result is a set of whole-word models similar to those implemented for standard keyword-filler word spotting systems. Thus, the individual states can be thought of as representing maximally context-dependent phonetic units, as they characterize a phonetic segment embedded in a particular context of the given word type.

2.2. Clustering Subword States Across Words

The above-defined HMM-training procedure produces a set $\mathcal{S}$ of word context-dependent subword states (cardinality $S = \sum_{w \in \mathcal{W}} Q_w$), each modeled with a Gaussian mixture model. The desired goal is to reduce this relatively high number S to a more manageable number K of less context-dependent states by clustering elements of $\mathcal{S}$ according to the pairwise similarities of their emission densities. One obvious approach would be to measure the density similarities directly by computing their Kullback-Leibler (KL) divergences. However, in the case of mixture models this cannot be computed analytically and would require a computationally intensive numerical approximation.

To circumvent this complication, we substitute density similarity with a simple measure of the correlation of state posterior trajectories as applied to our speech collection. In particular, if $X = x_1 x_2 \ldots x_T$ represents an arbitrary sample of speech, we compute a standard posteriorgram representation according to

$$P(s|x_t) = \frac{P(x_t|s)P(s)}{\sum_{s' \in \mathcal{S}} P(x_t|s')P(s')}, \tag{1}$$

where $P(x_t|s)$ is the standard GMM likelihood of observation x_t for state $s \in \mathcal{S}$. In practice, we assume uniform state priors, allowing the unknown $P(s)$ terms to drop out. We then compute the state similarity between $s, s' \in \mathcal{S}$ as the normalized inner product between the posterior trajectories given by

$$\mathrm{sim}(s, s') = \frac{\sum_{t=1}^{T} P(s|x_t)P(s'|x_t)}{\left[\sum_{t=1}^{T} P(s|x_t)\right]\left[\sum_{t=1}^{T} P(s'|x_t)\right]}. \tag{2}$$

This measure can be interpreted as posteriorgram cross correlation adapted from traditional form to this probabilistic setting.

We can use this state similarity measure to define a weighted undirected graph with one node per state. The edge weights are specified by the matrix $W_{ij} = \mathrm{sim}(s_i, s_j)$, where s_i and s_j are the states in $\mathcal{S}$ corresponding to nodes i and j in the graph, respectively. Given this graph and a desired number of clusters K, we can proceed with the spectral clustering variant defined in [11] as follows:

1. Compute the unnormalized graph Laplacian $L = D - W$, where D is the diagonal matrix with elements $D_{ii} = \sum_j W_{ij}$, the degree of the i-th vertex.

2. Solve the generalized eigenvalue problem $Lv = \lambda Dv$, for the first K eigenvectors $\{v_1, \ldots, v_K\}$, where each $v_i \in \mathbb{R}^S$.
3. Representing the i-th vertex (and thus the i-th state) by its graph spectrum $y_i = \langle v_1[i], v_2[i], \ldots, v_K[i]\rangle \in \mathbb{R}^K$, perform K-means clustering of the points $\{y_1, y_2, \ldots, y_S\}$.

The resulting collection of state clusters $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$, $c_i \subseteq \mathcal{S}$, can be used to transform the posteriorgram over states $s \in \mathcal{S}$ to a lower K-dimensional posteriorgram over clusters in $\mathcal{C}$ by $P(c_i|x_t) = \sum_{s \in c_i} P(s|x_t)$, where $P(s|x_t)$ is computed according to Eq. 1. Conceptually, the clusters in $\mathcal{C}$ define subword unit equivalence classes that will ideally produce posteriorgrams over some phonologically meaningful, speaker-independent categories. Below, we will refer to these K-dimensional posteriorgrams over state clusters as *unsupervised posteriorgrams*. Fig. 3 demonstrates that supervised and unsupervised posteriorgrams ($K = 100$) are remarkably similar.

3. Experiments

3.1. Evaluation Method

A suitable evaluation must determine how well a vector time series representation of speech can associate examples of the same word type spoken by a range of speakers while simultaneously preventing incorrect associations across word types. Using time aligned word transcripts for the Switchboard-1 corpus, we extracted all word examples that were at least 0.5 s in duration and at least 5 characters long as text, amounting to about 150k examples distributed over approximately 18k word types. It is from this set that we construct our word clusters for training by (i) taking $\mathcal{W}$ to be a set of some number of the most common word types, and (ii) defining the training clusters as a fraction f of the total examples available in the 150k set for each type. Reducing f simulates a reduction in term discovery efficiency.

From the 150k word set, we defined an 11k example subset for evaluation. For each of the $\binom{11k}{2} \approx 60$ million pairs of word examples in the subset, we computed the dynamic time warping (DTW) distance between them, using either cosine distance or symmetrized KL divergence as the frame-level distance metric. Each of the 60M DTW distances is between words of either the same of different type. Thus, we can view the isolation of correct from incorrect word matches as a retrieval task, sampling a precision-recall curve across the full range of DTW distance thresholds. We characterize the quality of each representation by the average precision. Note that approximately 100k of the 60M pairs are the same word type, while only 3k of those are the same word type spoken by the same person, making the average precision primarily a measure of speaker independence.

To evaluate our proposed method, we compare five baseline feature sets: PLP (39-dim including velocity and acceleration), PLP with principal component analysis (PCA) without reducing dimension, and phonetic posteriorgrams generated from supervised multilayer perceptron (MLP)-based acoustic models for English, German, and Spanish. The PLP and PLP+PCA features are computed directly from the acoustics and have no language specific knowledge. The English posteriorgrams, produced by MLPs trained on 100 hours of telephone speech data (see [10]), are highly supervised and set a sort of performance ceiling for the task. Finally, the evaluation of Spanish and German posteriorgrams, each produced by MLPs trained on 15 hours of telephone speech, provide a measure of how well an acoustic model in one language can characterize word-level similarities in another. Note that all three MLPs take PLP+PCA as input, as does our unsupervised training procedure.

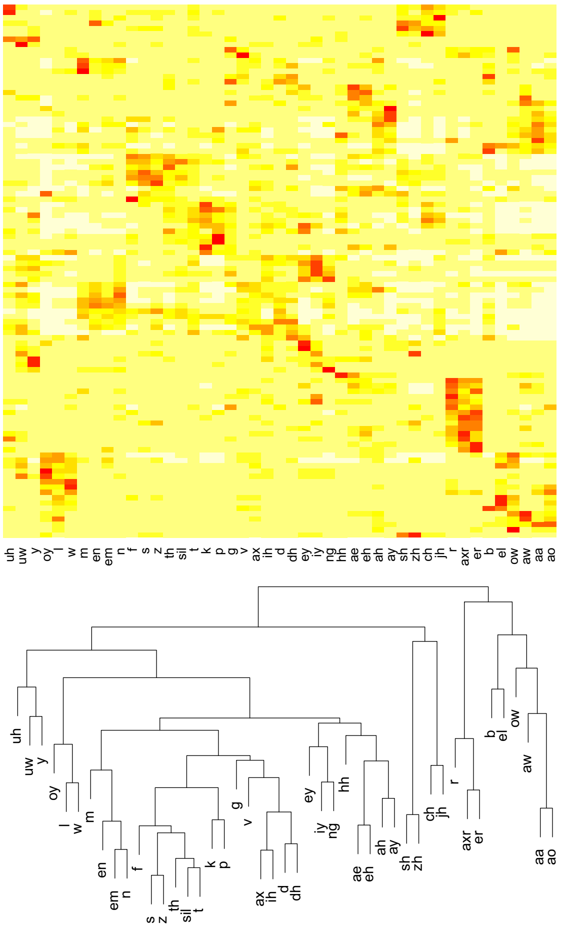

Figure 3: *Similar phones map to similar units. The mass in the heatmap (top) is concentrated in relatively small regions, associated with relatively few units (rows) and relatively few phones (columns). The dendrogram (bottom) shows the column labels in more detail. Note that similar phones are clustered near one another. These figures compare the proposed unsupervised units to supervised English posteriorgrams, using the similarity metric in Eq. 2.*

3.2. Results

Table 1 lists the average precision for the baseline features and various versions of the unsupervised posteriorgrams. For each feature type, we list the frame-level distance metric and the dimension of the resulting features. For the unsupervised posteriorgram features, we consider various settings for the fraction of word examples of each type used for model training (f), the number of word clusters/models ($|\mathcal{W}|$), and whether the state counts for each word (Q_w) were estimated (Est) or oracle (Ora). Note that for $f = 0.2$ we require each word cluster have at least 50 examples for training, which limited the number of types to 215; for $f = 0.2$ and $f = 0.1$ we require at least 20 examples, allowing 215 and 82 types, respectively. Also, the state number estimates use the regression of Fig. 2, which was performed using a separate collection of words from the model training set. State similarities of Eq. 2 were computed using the *evaluation* set (11k word examples) only. Through experimentation, we found that cosine distance was optimal for acoustic features and unsupervised posteriorgrams, while the MLP posteriorgrams, which tend toward low entropy, were best served by symmetrized KL divergence.

There are several trends apparent in the data. First, the matched English acoustic model far surpasses the Spanish and

Table 1: *Average precision (AP) performance on the word matching task for the baselines and the unsupervised posteriorgrams, considering multiple word clustering efficiencies (f) and estimated vs. oracle HMM state counts (SC). Also included are the number of word clusters ($|\mathcal{W}|$), the initial number of subword states ($S = \sum_w Q_w$), and the final dimension of the feature vectors (K for unsupervised posteriors).*

Features	f	$\|\mathcal{W}\|$	S	SC	Dim	AP
PLP	–	–	–	–	39	0.118
PLP+PCA	–	–	–	–	39	**0.169**
Spanish MLP	–	–	–	–	28	0.080
German MLP	–	–	–	–	46	**0.167**
Unsupervised Posteriorgrams	0.1	82	584	Ora	100	0.211
	0.2	215	1562	Ora	100	0.229
	0.5	215	1562	Ora	100	**0.290**
	0.1	82	582	Est	100	0.218
	0.2	215	1555	Est	100	0.247
	0.5	215	1535	Est	100	**0.312**
English MLP	–	–	–	–	45	**0.516**

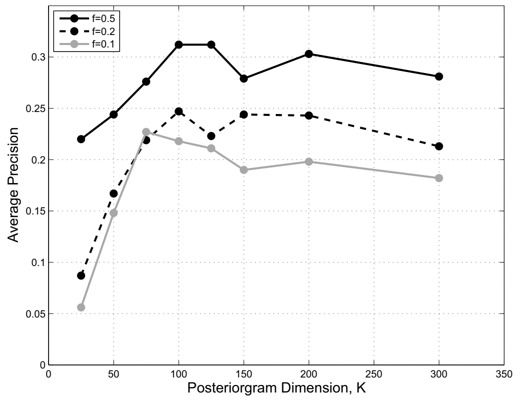

Figure 4: *Average precision as a function of the unsupervised posteriorgram dimension K for various word clustering efficiencies f.*

German acoustics applied to English speech. While this may not come as a surprise, it does indicate that the MLP's knowledge of speaker independence is highly language-specific, and thus the universality assumptions indicated in [10] have only limited validity. In fact, the word matching performance of the PLP features with PCA matched or exceeded both mismatched-language acoustic models. Second, the unsupervised posteriorgram performance for each of the training conditions significantly outperforms both acoustic features and mismatched MLP posteriorgrams. This demonstrates the potential for term discovery on constraining unsupervised acoustic model training to improve speaker independence in a zero resource setting. Finally, there is no loss in performance replacing the dictionary-provided number of subword states with our estimates based on median duration, even though the estimates can significantly deviate from the dictionary predicted number of phonetic states. This is likely a result of pronunciation dictionary inaccuracies for the conversational speaking style and a tolerance of our approach to deal with subword units that do not perfectly correspond to phones.

Fig. 4 plots the average precision of the unsupervised posteriorgram versus the number K of subword unit clusters for three values of the simulated clustering efficiency f. We find that the maximum average precision is achieved within the range of $K = 75$ to 125 for all three efficiencies, which amounts to approximately 1.5–2 times the number of English phones. This implies that under the state clustering procedure, preserving some level of context-dependency is optimal (see also Fig. 3). However, the optimal level of context dependency increases somewhat with increased amounts of word cluster training data, consistent with conventional wisdom in the supervised setting.

3.3. Remaining Challenges

The above results indicated that the performance increases with (i) more word clusters and (ii) more examples of each type. Moreover, in further experimentation we found word matching is improved for word types present in the training data, as the resulting subword models were more consistent for those cases. These facts imply that we can expect gains in performance as the amount of untranscribed training data increases; indeed, twice the data will mean twice the size of the word clusters, providing a more diverse set of words types with sufficient examples for model training. Therefore, the success of the proposed methods will rest heavily on the scaling capacity of the spoken term discovery algorithm, a component that we have taken for granted in this study. In particular, even if the word clustering it provides is relatively low efficiency, we can still produce an arbitrarily large number of word clusters for training if the discovery algorithm is efficient enough to search increasingly vast amounts of untranscribed speech for word-level matches. Given the inherent $O(n^2)$ nature of the term discovery task, scaling the search to hundreds or even thousands of hours will require significant research effort.

4. Conclusions

In the absence of word transcripts and pronunciation dictionary, we have presented a novel strategy for training a subword acoustic model using top-down speaker independence constraints. Despite remaining challenges in developing scalable spoken term discovery algorithms, we have demonstrated that the unlabeled word clusters they produce open the door for the unsupervised training of acoustic models with strong speaker independence properties.

5. References

[1] R. Singh, B. Raj, and R. M. Stern, "Automatic generation of subword units for speech recognition systems," *IEEE T-ASLP*, vol. 10, no. 2, pp. 89–99, 2002.

[2] A. Garcia and H. Gish, "Keyword spotting of arbitrary words using minimal speech resources," in *Proc. of ICASSP*, 2006.

[3] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised learning of acoustic subword units," in *ACL-08: HLT*, 2008.

[4] M.-H. Siu, H. Gish, A. Chan, and W. Belfield, "Improved topic classification and keyword discovery using an HMM-based speech recognizer trained without supervision," in *Proc. of Interspeech*, 2010.

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

[6] M. Bacchiani and M. Ostendorf, "Joint lexicon, acoustic unit inventory and model design," *Speech Communication*, vol. 29, pp. 99–114, 1999.

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

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

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

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

[11] J. Shi and J. Malik, "Normalized cuts and image segmenation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.