ABSTRACT

We study spoken term detection (STD) – the task of determining whether and where a given word or phrase appears in a given segment of speech – using articulatory feature-based pronunciation models. The models are motivated by the requirements of STD in low-resource settings, in which it may not be feasible to train a large-vocabulary continuous speech recognition system, as well as by the need to address pronunciation variation in conversational speech. Our STD system is trained to maximize the expected area under the receiver operating characteristic curve, often used to evaluate STD performance. In experimental evaluations on the Switchboard corpus, we find that our approach outperforms a baseline HMM-based system across a number of training set sizes, as well as a discriminative phone-based model in some settings.

Index Terms— spoken term detection, articulatory features, AUC, structural SVM, discriminative training

1. INTRODUCTION

Spoken term detection (STD) is the problem of determining whether, and optionally where, a given utterance contains a query term (a word or phrase) of interest. Typical STD approaches rely on large-vocabulary continuous speech recognition (LVCSR) systems trained on large amounts of data ([1, 2], inter alia). Such approaches are infeasible in low-resource settings, e.g. for languages or domains where training data are limited. In recent work, we have shown that a discriminative approach for STD can outperform a comparable HMM-based system in a limited-data setting [3].

In the current work, we explore an articulatory feature-based (AF-based) model for STD in conversational speech. Pronunciation variation in conversational speech is one of the leading causes of speech recognition errors [4, 5, 6]. Standard phone-based pronunciation models, which assume that phonemes are strung together to produce word pronunciations, have well-known drawbacks [7, 8]. Articulatory feature-based models (sometimes referred to as “production models”, “phonological feature models” or “gestural models” in the literature) have been proposed as an alternative [9, 10, 11, 12]; there is evidence that such approaches may improve recognition of noisy speech [13, 14, 15], adapt better across languages [16], improve hyperarticulated speech recognition [17], and address pronunciation variation [18, 19]. Some work has also begun to address discriminative training of AF-based models [20, 21].

Besides the potential benefit of articulatory models for conversational speech, it has also been argued that they should have advantages in low-resource settings due to their parsimony [7, 9]: While a given small training set may not contain sufficient examples of every context-dependent phone (or even monophone) to learn a robust model, many phones share the same articulatory features, so that articulatory models facilitate data sharing across phones. This work is therefore motivated by the needs of STD in both conversational speech settings and low-resource settings.

2. ARTICULATORY FEATURE-BASED MODEL

We address the pronunciation variation observed in conversational speech, as well as the challenges of a low-resource setting, with STD systems using an articulatory feature-based model, based on previous work by ourselves and others [10, 18, 22, 23]. The proposed model employs articulatory features that are based on the tract variables of articulatory phonology [24]. These variables represent the configurations of the speech articulators: the constriction degrees and positions of the lips, tongue tip, and tongue body; the state of the velum; and the state of the glottis. We build an AF-based baseform dictionary of canonical pronunciations by mapping the phones in a standard dictionary to their corresponding AF targets, expanding from the mapping defined in [25] to ensure a unique AF configuration for each phone.

We model pronunciation variation by allowing AF streams to transition asynchronously from one target state to the next. When all AFs are synchronized, the resulting surface pronunciation is identical to the canonical pronunciation; asynchronous transitions result in non-canonical pronunciations. Examples of non-canonical pronunciations resulting from asynchrony include nasalization, anticipatory/preservatory...
3. DISCRIMINATIVE MODEL FOR STD

We now turn to constructing a spoken term detector and training it using a discriminative algorithm based on [26]. Our goal is to learn a function \( f : \mathcal{X}^* \times \mathcal{Y}^* \to \mathbb{R} \), which takes as its input a speech utterance \( \mathbf{x} \in \mathcal{X}^* \) and a query term \( \mathbf{v} \in \mathcal{Y}^* \), where \( \mathcal{V} \) is the vocabulary of words, and returns a score \( f(\mathbf{x}, \mathbf{v}) \in \mathbb{R} \) representing the confidence that the query term occurs in the utterance. In a practical system, the utterance \( \mathbf{x} \) is declared to be a putative hit for a query term \( \mathbf{v} \) if \( f(\mathbf{x}, \mathbf{v}) > b \) for some threshold \( b \in \mathbb{R} \). We model the STD function, parameterized by a set of linear weights \( \mathbf{w} \in \mathbb{R}^n \), as

\[
f_w(\mathbf{x}, \mathbf{v}) = \max_{\mathbf{s} \in \mathcal{S}} \mathbf{w} \cdot \phi(\mathbf{x}, \mathbf{v}, \mathbf{s})
\]

where \( \mathcal{S} \) is the set of all valid articulatory segmentations and \( \phi(\mathbf{x}, \mathbf{v}, \mathbf{s}) \in \mathbb{R}^n \) is a feature vector. The score in Eq. 1 corresponds to the score of the highest scoring segmentation, \( \mathbf{s} \), over all possible start and end times within the utterance \( \mathbf{x} \) for the term \( \mathbf{v} \). The feature vectors, \( \phi(\mathbf{x}, \mathbf{v}, \mathbf{s}) \), are composed of a set of pre-defined feature maps \( \{\phi_j\}_{j=1}^M \), where \( \phi_j : \mathcal{X}^* \times \mathcal{Y}^* \times \mathcal{S} \to \mathbb{R}^r \). Each feature map takes as input the acoustics \( \mathbf{x} \), the term \( \mathbf{v} \), and the articulatory segmentation \( \mathbf{s} \) and returns an \( r \)-dimensional vector. We note that although the maximization in Eq. 1 is over an exponential number of possible segmentations, in the case where the feature maps are decomposable, the maximizing segmentation can be computed using dynamic programming as described in [23].

3.1. Feature Maps

We use two types of feature maps analogous to those used in our previous work on phone-based STD [3]. Our feature maps are constructed from a set of feature functions \( \bm{\xi} : \mathcal{X} \to \mathbb{R}^r \) computed from the acoustics \( \mathbf{x} \). The use of arbitrary feature functions allows us to leverage diverse sources of information. Given a suitable feature function \( \bm{\xi}(\cdot) \), our first set of feature maps compute the confidence that the acoustic frames correspond to the hypothesized configurations of AFs:

\[
\phi_1(\mathbf{x}, q^1, \ldots, q^K) = \frac{1}{s-e+1} \sum_{t=s}^{e} \bm{\xi}(\mathbf{x}_t) \delta[p_l(\mathbf{s}) = q^1] \cdots \delta[p_r(\mathbf{s}) = q^K]
\]

(2)

where each \( q^i \in Q_i \) is a value that AF stream \( i \) can take and \( \delta[\cdot] = 1 \) if the condition \( \cdot \) is true and 0 otherwise. Thus, we have \( |Q_1| \times \cdots \times |Q_K| \) feature maps of the first type, each of which is a vector of length equal to the length of \( \bm{\xi} \).

The second set of feature maps correspond to AF state transitions, and measure the relationship between the acoustics at a transition and its left/right states:

\[
\phi_2(\mathbf{x}, q_1, q_2) = \frac{1}{s-e+1} \sum_{t=s+1}^{e} \bm{\xi}(\mathbf{x}_t) \delta[p_l(\mathbf{s}) = q_1] \delta[p_r(\mathbf{s}) = q_2]
\]

(3)

where \( q_1, q_2 \in Q^i \) are possible states for stream \( i \). As in Eq. 2, each feature map is a vector of length equal to the length of \( \bm{\xi} \) with a total of \( \sum_{i=1}^{K} |Q_i|^2 \) feature maps of this type.

Note that the feature maps in Eqs. 2 and 3 are normalized

---

1 Another component of pronunciation variation, besides asynchrony, is substitution (typically reduction) of one AF value for another. This has been explored in other work (e.g., [19]) and is not modeled explicitly here, but it is implicitly modeled by AF classifier posteriors; see subsequent sections.

2 In experiments, we assume that lip features form a fully synchronized “bundle”, as do all tongue features and the pair (glottis, velum), so \( K = 3 \).
by the length of region in which the term has been hypothesized, in order to make scores comparable across different segment lengths. Also, we note that if we restrict the model to contain only a single stream, whose values correspond to the phoneme sequence in the term’s pronunciation, then the resulting feature maps are identical to those used in our previous STD approach using a phone-based model [3].

3.2. Large-Margin Training to Optimize AUC

The STD function defined in Eq. 1 represents the confidence that the term \( \tau \) was uttered in the utterance \( X \). For a given threshold \( b \), the utterance is declared to contain the term if \( f_w(X, \tau) > b \). The trade-off between hits and misses can be quantified using the receiver operating characteristic (ROC) curve, which is the true-positive (detection) rate versus false-negative rate across the range of possible thresholds. The area under the ROC curve (AUC) is a measure of performance averaged across all possible thresholds, which ranges from 0.5 (chance performance) to 1 (perfect detection). Our goal is to learn the model parameters \( w \) in Eq. 1 so as to maximize the AUC on unseen data. We do this using the algorithm described in [3], which we briefly outline here for completeness.

We assume that we can construct a set of \( N \) training examples \( T = \{ \tau_i, X_i^+, X_i^-, s_i^+, e_i^+ \}_{i=1}^N \), where each example consists of a query term \( \tau_i \in \mathcal{V} \), a “positive” utterance \( X_i^+ \) that contains the term, a “negative” utterance \( X_i^- \) in which the term is absent, and the start and end frames of the term in the positive utterance \( (s_i^+, e_i^+) \). The configuration of weights that maximizes the expected AUC is related to the Wilcoxon-Mann-Whitney statistic [27]. We determine the optimal set of weights by minimizing the following regularized structural hinge loss over the training set:

\[
\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{\lambda}{2} |\mathbf{w}|^2 + \frac{1}{N} \sum_{i=1}^N \left[ 1 - f_w(\mathbf{X}_i^+, \tau_i) + f_w(\mathbf{X}_i^-, \tau_i) \right]_+ \tag{4}
\]

where \( |x|_+ = \max\{0, x\} \) and \( \lambda \) is a regularization parameter that prevents overfitting. Note that we require that the location of the query term in the positive utterance \( X_i^+ \) be known, but we do not require knowledge of the segmentation \( \mathbf{s}_i^+ \). Therefore, unlike the algorithm in [26], our algorithm can be applied without having to first compute an articulatory forced alignment for the utterance. In computing \( f_w(\mathbf{X}_i^+, \tau_i) \) we restrict the search to only those segmentations \( \mathbf{s} \) that begin and end at the appropriate times: \( f_w(\mathbf{X}_i^+, \tau_i) = \max_{\mathbf{s} \sim (s_i^+, e_i^+)} \mathbf{w} \cdot \phi(\mathbf{X}_i^+, \tau_i, \mathbf{s}) \). In computing \( f_w(\mathbf{X}_i^-, \tau_i) \), we search over all possible start and end times.

For additional algorithmic details, including pseudocode, see [3]. Note that this approach differs significantly from other recent work on discriminatively trained AF-based models since our models are applied to a prediction task involving acoustics (as opposed to lexical access as in [20, 21]).

4. EXPERIMENTS

We conduct experiments on the Switchboard corpus [28] of conversational speech. To facilitate comparison with our previous phone-based STD work, we use the same experimental setup as in [3]. We compare performance obtained by training on four sets of increasing size containing 500, 1000, 2500, and 5000 utterances selected from Switchboard sets 23-49; each larger set contains all utterances from the smaller set. A development set of 40 terms is used for parameter tuning, and results are reported on a test set containing 60 terms. For each term in the development and test sets, we consider 20 utterances containing the term (positive utterances) and 20 utterances that do not contain the term (negative utterances), drawn from Switchboard sets 20-22. Initial and final silences are removed from all utterances.\(^3\)

To define the training set, we begin by identifying each instance of a word containing at least five phonemes in its canonical pronunciation as a candidate term \( \tau \) and considering the corresponding utterance as a positive example for that term \( X_i^+ \). We randomly select an utterance \( X_i^- \) that does not contain \( \tau \) to serve as a corresponding negative example for the term. The chosen training pairs are identical to those used in the experiments reported in [3].

Following [3], we parameterize the acoustics using 12th order PLP coefficients with energy, deltas, and double-deltas to obtain a 39-dimensional input representation \((X \subseteq \mathbb{R}^{39})\). We train four multi-layer perceptrons (MLPs): three to predict lip state (L, 5 labels), tongue state (T, 25 labels), and glottis-velum state (G, 10 labels); and one to predict phone labels. We train the MLPs on all of the transcribed STP data corresponding to Switchboard sets 23-49 using the QuickNet toolkit [29]. We concatenate each frame of MLP coefficients with the four preceding and succeeding frames to form a 351-dimensional input representation for the MLPs. The MLPs are single hidden layer feed-forward networks trained to optimize a cross-entropy criterion, with the number of hidden layer nodes determined by tuning on a held-out portion of the training data. Once the MLPs have been trained, we compute log-posteriors from the nets, concatenate them, and project the resulting features onto the top 39 principal components to obtain a tandem feature representation [30] that forms the feature functions \( \xi(x) \), to which we append a constant bias term (so that \(|\xi(x)| = 40\)). These feature functions are used in our discriminative STD systems and as acoustic features in our GMM-HMM baselines.

We compare against two HMM-based baseline systems trained using HTK [31]. Our baselines are constructed by concatenating 3-state HMM models representing the query term phones in parallel with a garbage model containing every other phone model. We consider both a context-independent monophone baseline (HMM-mono) and a context-dependent

\(^3\)Details of the utterances and query terms used in these experiments can be found at http://www.ttic.edu/keshet/Keyword_Spotting.html.
Table 1. AUC averaged over 60 query terms in the test set for systems trained on 500-5000 utterances. *, † = significant (p ≤ 0.05) improvement over HMM-tri and Disc-Phone, respectively, using a 1-tailed Wilcoxon signed-ranks test.

<table>
<thead>
<tr>
<th>System</th>
<th>500</th>
<th>1000</th>
<th>2500</th>
<th>5000</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM-mono</td>
<td>0.810</td>
<td>0.827</td>
<td>0.846</td>
<td>0.857</td>
</tr>
<tr>
<td>HMM-tri</td>
<td>0.828</td>
<td>0.855</td>
<td>0.899</td>
<td>0.920</td>
</tr>
<tr>
<td>Disc-Phone [23]</td>
<td>0.874*</td>
<td>0.901*</td>
<td>0.917</td>
<td>0.933*</td>
</tr>
<tr>
<td>Disc-AF-0</td>
<td>0.885*†</td>
<td>0.897*†</td>
<td>0.913</td>
<td>0.936*†</td>
</tr>
<tr>
<td>Disc-AF-1</td>
<td>0.888*†</td>
<td>0.898*†</td>
<td>0.915</td>
<td>0.939*†</td>
</tr>
<tr>
<td>Disc-Phone-AF-1</td>
<td>0.891*†</td>
<td>0.905*†</td>
<td>0.920*†</td>
<td>0.940*†</td>
</tr>
</tbody>
</table>

5. DISCUSSION AND ANALYSIS

All of the discriminative systems significantly outperform the monophone HMM baseline. For all training set sizes except 2500, the discriminative systems also outperform the context-dependent HMM baseline. This is particularly encouraging, because our discriminative systems are context-independent. It is fairly straightforward to add context dependence to our discriminative models; we leave this as future work.

The AF-based systems significantly outperform the phone-based discriminative system in the lowest-data case (p < 0.025). In the highest data case, the difference in AUC is not significant at p = 0.033. The AF-based system with asynchrony (Disc-AF-1; M = 1) performs better than the synchronous system (Disc-AF-0; M = 0), and assigning 3 states per AF label. We note that the system with no asynchrony is not identical to a discriminative phone-based system (as in [3]), because of the different feature maps. Our results are summarized in Table 1.

We compare the baseline systems against our AF-based discriminative systems allowing either one state of asynchrony (Disc-AF-1; M = 1) or no asynchrony (Disc-AF-0; M = 0), and assigning 3 states per AF label. We note that the system with no asynchrony is not identical to a discriminative phone-based system (as in [3]), because of the different feature maps. Our results are summarized in Table 1.

In future work, we would like to incorporate context dependence in our models in order to further improve performance, to consider additional feature maps, to explore discriminative optimization of other criteria such as the figure of merit (FOM) [32] or actual term-weighted value (ATWV) [33], and to test on low-resource languages. While the approach is intended for low-resource settings, trained discriminatively to optimize a task-specific criterion. In experiments on low-resource Switchboard STD, the proposed system outperforms our previous phone-based STD system [3] in the lowest and highest data setting, outperforms a context-dependent HMM baseline across multiple training set sizes, and performs better still when combined with our discriminative phone-based model.

In future work, we would like to incorporate context dependence in our models in order to further improve performance, to consider additional feature maps, to explore discriminative optimization of other criteria such as the figure of merit (FOM) [32] or actual term-weighted value (ATWV) [33], and to test on low-resource languages. While the approach is intended for low-resource settings, trained discriminatively to optimize a task-specific criterion. In experiments on low-resource Switchboard STD, the proposed system outperforms our previous phone-based STD system [3] in the lowest and highest data setting, outperforms a context-dependent HMM baseline across multiple training set sizes, and performs better still when combined with our discriminative phone-based model.

6. CONCLUSIONS

We have presented an articulatory feature-based model for STD, motivated by the challenges of low-resource and conversational settings, trained discriminatively to optimize a task-specific criterion. In experiments on low-resource Switchboard STD, the proposed system outperforms our previous phone-based STD system [3] in the lowest and highest data setting, outperforms a context-dependent HMM baseline across multiple training set sizes, and performs better still when combined with our discriminative phone-based model.
7. REFERENCES


