DISCRIMINATIVE ARTICULATORY MODELS FOR SPOKEN TERM DETECTION IN LOW-RESOURCE CONVERSATIONAL SETTINGS

Rohit Prabhavalkar[1] *Karen Livescu*[2] *Eric Fosler-Lussier*[1] *Joseph Keshet*[3]

[1] The Ohio State University, USA; [2] TTI-Chicago, USA; [3] Bar-Ilan University, Israel
{prabhava, fosler}@cse.ohio-state.edu klivescu@ttic.edu jkeshet@cs.biu.ac.il

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 variability 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

We would like to thank Yanzhang He and Preethi Jyothi for helpful comments and suggestions. This research was supported by NSF grants IIS-0905420 and IIS-0905633. The opinions expressed in this work are those of the authors and do not necessarily reflect the views of the funding agency.

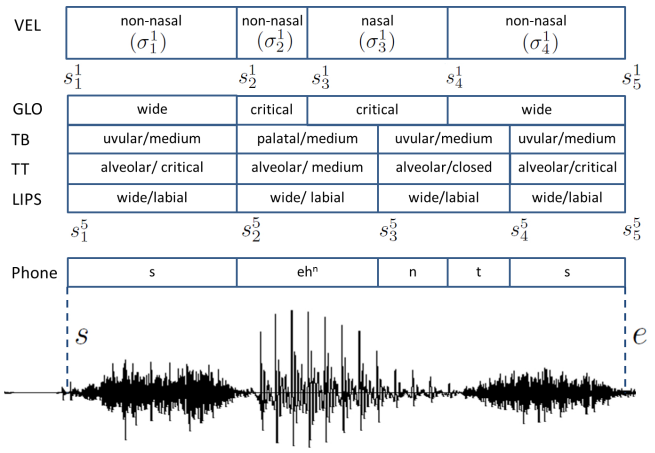

Fig. 1. Non-canonical pronunciation of the word 'sense'. The glottis and velum desynchronize from the other features, producing an epenthetic [t] and nasalized [eh].

rounding, and epenthetic stop insertions (see Fig. 1).[1]

Formally, we model pronunciation via a set of K articulatory feature streams.[2] We assume that the waveform is parameterized into acoustic feature vectors (e.g., PLPs) $\overline{\mathbf{x}} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T)$, where $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^d$ is a feature vector for frame t. Given an utterance $\overline{\mathbf{x}}$ and a query term $\overline{v}$, we denote by $|\overline{v}|$ the number of phones in the canonical pronunciation of $\overline{v}$. We denote the corresponding sequence of articulatory targets for stream i as $(\sigma_1^i, \sigma_2^i, \cdots, \sigma_{|\overline{v}|}^i)$. For a given hypothesized start and end time, $(1 \leq s < e \leq T)$, we denote a valid articulatory segmentation $\overline{\mathbf{s}}$ of $\overline{v}$ as the matrix of values that represent the start and end times for each of the AF states: $\overline{\mathbf{s}}_{i,j} = s_j^i$ where s_j^i is the start time of the j^{th} unit in stream i (i.e. σ_j^i). Thus, $s = s_1^i < s_2^i < \cdots < s_{|\overline{v}|}^i < e$, so that the state j in stream i extends from $t = s_j^i$ to $t = s_{j+1}^i - 1$, where $s_{|\overline{v}|+1}^i = e$. We use the notation $\overline{\mathbf{s}} \sim (s, e)$ to denote an articulatory segmentation $\overline{\mathbf{s}}$ that begins at frame s and ends at frame e.

In order to reduce computational complexity [23] and eliminate implausible segmentations, we restrict the amount of asynchrony to some number of states M: For all pairs of streams i, j and for each unit $1 \leq k \leq |\overline{v}|$ in the pronunciation, the extent of σ_k^i must lie between the extents of the succeeding and preceding M units in all other streams, i.e. $s_{k-M}^j \leq s_k^i$ and $s_{k+1}^i \leq s_{k+M}^j$. In particular, setting $M = 0$ would enforce complete synchrony. Finally, we denote the AF value for stream i hypothesized at time frame t under segmentation $\overline{\mathbf{s}}$ as $p_t^i(\overline{\mathbf{s}})$, i.e. $p_t^i(\overline{\mathbf{s}}) = \sigma_j^i$ for $s_j^i \leq t < s_{j+1}^i$. Our notation is presented in Fig. 1.

[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 $K = 3$.

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 $f : \mathcal{X}^* \times \mathcal{V}^* \rightarrow \mathbb{R}$, which takes as its input a speech utterance $\overline{\mathbf{x}} \in \mathcal{X}^*$ and a query term $\overline{v} \in \mathcal{V}^*$, where $\mathcal{V}$ is the vocabulary of words, and returns a score $f(\overline{\mathbf{x}}, \overline{v}) \in \mathbb{R}$ representing the confidence that the query term occurs in the utterance. In a practical system, the utterance $\overline{\mathbf{x}}$ is declared to be a putative hit for a query term $\overline{v}$ if $f(\overline{\mathbf{x}}, \overline{v}) > b$ 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

$$f_{\mathbf{w}}(\overline{\mathbf{x}}, \overline{v}) = \max_{\overline{\mathbf{s}} \in \mathcal{S}} \mathbf{w} \cdot \boldsymbol{\phi}(\overline{\mathbf{x}}, \overline{v}, \overline{\mathbf{s}}) \quad (1)$$

where $\mathcal{S}$ is the set of all valid articulatory segmentations and $\boldsymbol{\phi}(\overline{\mathbf{x}}, \overline{v}, \overline{\mathbf{s}}) \in \mathbb{R}^n$ is a feature vector. The score in Eq. 1 corresponds to the score of the highest scoring segmentation, $\overline{\mathbf{s}}$, over all possible start and end times within the utterance $\overline{\mathbf{x}}$ for the term $\overline{v}$. The feature vectors, $\boldsymbol{\phi}(\overline{\mathbf{x}}, \overline{v}, \overline{\mathbf{s}})$, are composed of a set of pre-defined feature maps $\{\phi_j\}_{j=1}^m$, where $\phi_j : \mathcal{X}^* \times \mathcal{V}^* \times \mathcal{S} \rightarrow \mathbb{R}^r$. Each feature map takes as input the acoustics $\overline{\mathbf{x}}$, the term $\overline{v}$, and the articulatory segmentation $\overline{\mathbf{s}}$ 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* $\boldsymbol{\xi} : \mathcal{X} \rightarrow \mathbb{R}^r$ computed from the acoustics $\overline{\mathbf{x}}$. The use of arbitrary feature functions allows us to leverage diverse sources of information. Given a suitable feature function $\boldsymbol{\xi}(\cdot)$, our first set of feature maps compute the confidence that the acoustic frames correspond to the hypothesized configurations of AFs:

$$\phi_{1,q^1,\cdots,q^K} = \frac{1}{s-e+1} \sum_{t=s}^{e} \boldsymbol{\xi}(\mathbf{x}_t) \delta_{[p_t^1(\overline{\mathbf{s}})=q^1]} \cdots \delta_{[p_t^K(\overline{\mathbf{s}})=q^K]} \quad (2)$$

where each $q^i \in \mathcal{Q}^i$ is a value that AF stream i can take and $\delta_{[a]} = 1$ if the condition a is true and 0 otherwise. Thus, we have $|\mathcal{Q}_1| \times \cdots \times |\mathcal{Q}_K|$ features maps of the first type, each of which is a vector of length equal to the length of $\boldsymbol{\xi}$.

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:

$$\phi_{2,i,q_1,q_2} = \frac{1}{s-e+1} \sum_{t=s+1}^{e} \boldsymbol{\xi}(\mathbf{x}_t) \delta_{[p_{t-1}^i(\overline{\mathbf{s}})=q_1]} \delta_{[p_t^i(\overline{\mathbf{s}})=q_2]} \quad (3)$$

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

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

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 $\overline{v}$ was uttered in the utterance $\overline{\mathbf{x}}$. For a given threshold b, the utterance is declared to contain the term if $f_{\mathbf{w}}(\overline{\mathbf{x}}, \overline{v}) > b$. 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-positive 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 $\mathbf{w}$ 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 $\mathcal{T} = \{\overline{v}_i, \overline{\mathbf{x}}_i^+, \overline{\mathbf{x}}_i^-, s_i^+, e_i^+\}_{i=1}^N$, where each example consists of a query term $\overline{v}_i \in \mathcal{V}^*$, a "positive" utterance $\overline{\mathbf{x}}_i^+$ that contains the term, a "negative" utterance $\overline{\mathbf{x}}_i^-$ 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:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{\lambda}{2} ||\mathbf{w}||^2 + \frac{1}{N} \sum_{i=1}^{N} [1 - f_{\mathbf{w}}(\overline{\mathbf{x}}_i^+, \overline{v}_i) + f_{\mathbf{w}}(\overline{\mathbf{x}}_i^-, \overline{v}_i)]_+ \tag{4}$$

where $[x]_+ = \max\{0, x\}$ and λ is a regularization parameter that prevents overfitting. Note that we require that the location of the query term in the positive utterance $\overline{\mathbf{x}}_i^+$ be known, but we do not require knowledge of the segmentation $\overline{\mathbf{s}}_i^+$. 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 $f_{\mathbf{w}}(\overline{\mathbf{x}}_i^+, \overline{v}_i)$ we restrict the search to only those segmentations $\overline{\mathbf{s}}$ that begin and end at the appropriate times: $f_{\mathbf{w}}(\overline{\mathbf{x}}_i^+, \overline{v}_i) = \max_{\overline{\mathbf{s}} \sim (s_i^+, e_i^+)} \mathbf{w} \cdot \boldsymbol{\phi}(\overline{\mathbf{x}}_i^+, \overline{v}_i, \overline{\mathbf{s}})$. In computing $f_{\mathbf{w}}(\overline{\mathbf{x}}_i^-, \overline{v}_i)$, 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 $\overline{v}_i$ and considering the corresponding utterance as a positive example for that term $\overline{\mathbf{x}}_i^+$. We randomly select an utterance $\overline{\mathbf{x}}_i^-$ that does not contain $\overline{v}_i$ 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 12^{th} order PLP coefficients with energy, deltas, and double-deltas to obtain a 39-dimensional input representation ($\mathcal{X} \subseteq \mathbb{R}^{39}$). 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 PLP 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 $\boldsymbol{\xi}(\overline{\mathbf{x}})$, to which we append a constant bias term (so that $|\boldsymbol{\xi}(\mathbf{x})| = 40$). 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.

System	500	1000	2500	5000
HMM-mono	0.810	0.827	0.846	0.857
HMM-tri	0.828	0.855	0.899	0.920
Disc-Phone [23]	0.874*	0.901*	0.917	0.933*
Disc-AF-0	0.885*,†	0.897*	0.914	0.937*
Disc-AF-1	0.888*,†	0.898*	0.915	0.939*,†
Disc-Phone-AF-1	0.891*,†	0.905*	0.920*	0.940*,†

Table 1. AUC averaged over 60 query terms in the test set for systems trained on 500-5000 utterances. *, † = significant ($p \leq 0.05$) improvement over HMM-tri and Disc-Phone, respectively, using a 1-tailed Wilcoxon signed-ranks test.

word-internal triphone baseline (HMM-tri). Output distributions are modeled as GMMs, with the number of mixture components determined by tuning on the development set. In both cases, we trade off between true positives and false positives by varying the term insertion probability. A term is detected if the 1-best Viterbi decoding hypothesis passes through the HMMs representing the term. By varying the term insertion probability, we can generate the ROC, and therefore the AUC, for each term. We also compare to a discriminative phone-based baseline system with 3 states per phone (Disc-Phone, referred to as SystemB in [3]).

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.

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 between Disc-AF-1 and Disc-Phone is at a significance level of $p = 0.033$. The AF-based system with asynchrony (Disc-AF-1) slightly outperforms the synchronous system (Disc-AF-0) across data set sizes, but the differences are insignificant. Combining phone and AF-based models (Disc-Phone-AF-1), i.e.

$$f_{\mathbf{w}}(\overline{\mathbf{x}}, \overline{v}) = \max_{\overline{\mathbf{s}}_{\mathrm{P}}, \overline{\mathbf{s}}_{\mathrm{AF}}} \mathbf{w}_{\mathrm{P}} \boldsymbol{\phi}_{\mathrm{P}}(\overline{\mathbf{x}}, \overline{v}, \overline{\mathbf{s}}_{\mathrm{P}}) + \mathbf{w}_{\mathrm{AF}} \boldsymbol{\phi}_{\mathrm{AF}}(\overline{\mathbf{x}}, \overline{v}, \overline{\mathbf{s}}_{\mathrm{AF}}) \quad (5)$$

where we constrain $\overline{\mathbf{s}}_{\mathrm{P}}$ and $\overline{\mathbf{s}}_{\mathrm{AF}}$ to have the same start and end times, and the weights $\mathbf{w}_{\mathrm{P}}$ and $\mathbf{w}_{\mathrm{AF}}$ are initialized using the trained models Disc-Phone and Disc-AF-1 and then trained discriminatively, improves further, significantly outperforming HMM-tri in every case and the discriminative

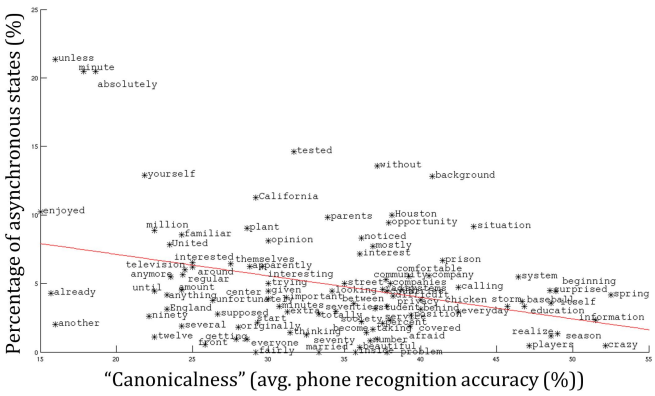

Fig. 2. Fraction of hypothesized asynchronous states vs. "canonicalness" of the pronunciation, for the 100 query terms in the development and test sets in the 5000-utterance condition. The line is the best linear fit to the data, showing an overall tendency to hypothesize more asynchrony for terms with less-canonical pronunciations.

phone-based system in the lowest- and highest-data cases.

We further analyze the behavior of the Disc-AF-1 system trained on 5000 utterances to understand when it hypothesizes asynchrony. We computed unconstrained phonetic decodings using the HMM-mono system on the portion of the positive utterances corresponding to the query term. The phonetic accuracies of these decodings against the canonical pronunciations give a rough measure of pronunciation variation in utterances of that term. We also examined the segmentations hypothesized by the AF-based system to determine the percentage of states that are asynchronous. Fig. 2 shows this percentage vs. the "canonicalness" measure for each keyword, and suggests that, as expected, the AF-based system is hypothesizing asynchrony for utterances with higher pronunciation variation.

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.

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, it would be interesting to compare performance to other previously published results on, e.g., NIST06 STDEVAL.

7. REFERENCES

[1] D. R. H. Miller, M. Kleber, C.-L. Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in *Proc. Interspeech*, 2007.

[2] D. Vergyri, I. Shafran, A. Stolcke, R. R. Gadde, M. Akbacak, B. Roark, and W. Wang, "The SRI/OGI 2006 spoken term detection system," in *Proc. Interspeech*, 2007.

[3] R. Prabhavalkar, J. Keshet, K. Livescu, and E. Fosler-Lussier, "Discriminative spoken term detection with limited data," in *Symposium on Machine Learning in Speech and Language Processing (MLSLP)*, 2012, Online: http://www.ttic.edu/sigml/symposium2012/papers/prabhavalkar_mlslp2012.pdf.

[4] D. McAllaster, L. Gillick, F. Scattone, and M. Newman, "Fabricating conversational speech data with acoustic models : A program to examine model-data mismatch," in *Proc. ICSLP*, 1998.

[5] M. Weintraub, K. Taussig, K. Hunicke-Smith, and A. Snodgrass, "Effect of speaking style on LVCSR performance," in *Proc. ICSLP*, 1996.

[6] K. Livescu, E. Fosler-Lussier, and F. Metze, "Subword modeling for automatic speech recognition: Past, present, and emerging approaches," *Signal Processing Magazine*, vol. 29, no. 6, pp. 44–57, 2012.

[7] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," in *Proc. ASRU*, 1999.

[8] D. Jurafsky, W. Ward, Z. Jianping, K. Herold, Y. Xiuyang, and Z. Sen, "What kind of pronunciation variation is hard for triphones to model?," in *Proc. ICASSP*, 2001.

[9] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[10] L. Deng, G. Ramsay, and D. Sun, "Production models as a structural basis for automatic speech recognition," *Speech Communication*, vol. 33, no. 2–3, pp. 93–111, 1997.

[11] S. King, T. Stephenson, S. Isard, P. Taylor, and A. Strachan, "Speech recognition via phonetically-featured syllables," in *Proc. Workshop on Phonetics and Phonology in ASR "Phonus 5"*, 2000.

[12] R. C. Rose, J. Schroeter, and M. M. Sondhi, "The potential role of speech production models in automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 99, no. 3, pp. 1699–1709, 1996.

[13] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, no. 3-4, pp. 303 – 319, 2002.

[14] K. Livescu, J. Glass, and J. Bilmes, "Hidden feature models for speech recognition using dynamic Bayesian networks," in *Proc. Eurospeech*, 2003.

[15] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Gesture-based dynamic Bayesian network for noise robust speech recognition," in *Proc. ICASSP*, 2011.

[16] S. Stüker, F. Metze, T. Schultz, and A. Waibel, "Integrating multilingual articulatory features into speech recognition," in *Proc. Eurospeech*, 2003.

[17] H. Soltau, F. Metze, and A. Waibel, "Compensating for Hyperarticulation by Modeling Articulatory Properties," in *Proc. ICSLP*, 2002.

[18] K. Livescu and J. Glass, "Feature-based pronunciation modeling with trainable asynchrony probabilities," in *Proc. ICSLP*, 2004.

[19] P. Jyothi, K. Livescu, and E. Fosler-Lussier, "Lexical access experiments with context-dependent articulatory feature-based models," in *Proc. ICASSP*, 2011.

[20] H. Tang, J. Keshet, and K. Livescu, "Discriminative pronunciation modeling: A large-margin, feature-rich approach," in *Proc. Association for Computational Linguistics (ACL)*, 2012.

[21] P. Jyothi, E. Fosler-Lussier, and K. Livescu, "Discriminatively learning factorized finite state pronunciation models from dynamic Bayesian networks," in *Proc. Interspeech*, 2012.

[22] K. Livescu, Ö. Çetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman, S. Dawson-Haggerty, B. Woods, J. Frankel, M. Magimai-Doss, and K. Saenko, "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop," in *Proc. ICASSP*, 2007.

[23] R. Prabhavalkar, E. Fosler-Lussier, and K. Livescu, "A factored conditional random field model for articulatory feature forced transcription," in *Proc. ASRU*, 2011.

[24] C. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, pp. 155–180, 1992.

[25] K. Livescu, *Feature-based Pronunciation Modeling for Automatic Speech Recognition*, Ph.D. thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2005.

[26] J. Keshet, D. Grangier, and S. Bengio, "Discriminative keyword spotting," *Speech Communication*, vol. 51, pp. 317–329, 2009.

[27] C. Cortes and M. Mohri, "Confidence intervals for the area under the ROC curve," in *Proc. NIPS*, 2004.

[28] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in *Proc. ICASSP*, 1992.

[29] D. Johnson et al., "ICSI QuickNet software package," http://www.icsi.berkeley.edu/Speech/qn.html, 2004.

[30] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, 2000.

[31] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University Press, 2002.

[32] R. Wallace, B. Baker, R. Vogt, and S. Sridharan, "Discriminative optimisation of the figure of merit for phonetic spoken term detection," *IEEE. Trans. Audio, Speech, and Language Processing*, vol. 19, pp. 1677–1687, 2011.

[33] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddington, "Results of the 2006 spoken term detection evaluation," in *Proc. ACM SIGIR Workshop on Searching Spontaneous Conversational Speech*, 2007.