DISCRIMINATIVE SPOKEN TERM DETECTION WITH LIMITED DATA

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

[1]Department of Computer Science and Engineering, The Ohio State University
[2]TTI-Chicago

{prabhava,fosler}@cse.ohio-state.edu {jkeshet,klivescu}@ttic.edu

ABSTRACT

We study spoken term detection—the task of determining whether and where a given word or phrase appears in a given segment of speech—in the setting of limited training data. This setting is becoming increasingly important as interest grows in porting spoken term detection to multiple low-resource languages and acoustic environments. We propose a discriminative algorithm that aims at maximizing the area under the receiver operating characteristic curve, often used to evaluate the performance of spoken term detection systems. We implement the approach using a set of feature functions based on multilayer perceptron classifiers of phones and articulatory features, and experiment on data drawn from the Switchboard database of conversational telephone speech. Our approach outperforms a baseline HMM-based system by a large margin across a number of training set sizes.

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

1. INTRODUCTION

Spoken term detection (STD) is the problem of determining whether and where a target word or multi-word phrase has been uttered in a speech recording. Many STD systems are based on large-vocabulary automatic speech recognition (ASR) systems, trained on very large amounts of data (e.g., [1]); in such systems, the STD task becomes the problem of searching a speech recording that has already been recognized and indexed using a large ASR system.

However, it is not always appropriate to assume that large amounts of training data will be available: rapid development of STD systems for low-resource languages, or porting STD systems to new acoustic conditions or speech styles are two example scenarios where data may be limited. In this work, we study the problem of developing STD systems that utilize limited data. By developing a discriminative approach that optimizes STD performance (as measured by the area under the receiver operating characteristic (ROC) curve, or AUC), we can construct systems that outperform ASR-based systems that are optimized for speech transcription performance.

Optimizing the expected AUC is accomplished by maximizing a weighted sum of feature functions for all possible locations of the query term in the input signal. The actual expectation cannot be computed, as we do not know the distribution over terms and utterances that examples are drawn from. One proxy would be to estimate the weights from a regularized average of the AUC over the training set, but this leads to an intractable combinatorial optimization problem. Instead we learn the weights by maximizing a convex lower bound of the AUC.

The linear discriminative approach allows easy integration of rich feature functions. In this work, however, we use a fairly limited set of feature functions that allow us to make direct comparisons with baseline HMM-based systems, and defer the use of more complex feature functions to future work.

We next present the model and learning algorithm, and describe experiments on detection of single-word terms in utterances from the Switchboard corpus. Our approach builds on previous work on discriminative keyword spotting [2], adapting both the learning algorithm to handle unknown segmental time alignments, and feature functions to be similar to those used in tandem-based HMM systems. Adapting the algorithm to handle unknown time alignments allows it to be deployed in a limited data setting where it may not be possible to robustly estimate the alignments using a forced aligner.

2. PROBLEM SETTING

In this section, we formally describe our problem setting and notation, depicted in Fig. 1. The speech signal is composed of a sequence of T acoustic feature vectors $\overline{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$, where $\mathbf{x}_t \in \mathcal{X} \subset \mathbb{R}^d$, $1 \leq t \leq T$, is a d-dimensional feature vector extracted from the t^{th} frame. We denote a term by $\overline{v} \in \mathcal{V}^*$, where $\mathcal{V}^*$ denotes a sequence of one or more words from a lexicon $\mathcal{V}$. We assume that we have a pronunciation dictionary $\pi : \mathcal{V} \rightarrow \mathcal{P}^*$ that given a word v provides us with its canonical pronunciation $\overline{p}^v = (p_1, \ldots, p_{L^v}) \in \mathcal{P}^*$, where $\mathcal{P}^*$ is the set of all finite-length phone sequences over the phone set $\mathcal{P}$ and L^v is the number of phones in the word's pronunciation. For example, for $v =$ "sense", its canonical pronunciation is $\overline{p}^v = (s, eh, n, s)$, with $L^v = 4$. The pronunciation of a term $\overline{v}$, which is a sequence of words, is denoted $\overline{p}^{\overline{v}}$ and is the concatenation of the pronunciations of the words composing $\overline{v}$. We define a sequence of phone start times and

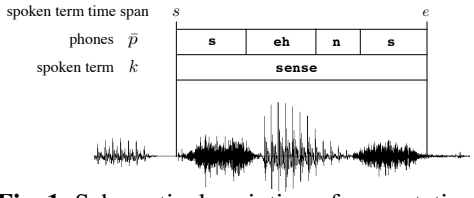

Fig. 1. Schematic description of our notation.

an end time for the pronunciation $\overline{p}^{\overline{v}}$ as $\overline{s} = (s_1, \ldots, s_{L^{\overline{v}}}, e)$, where s_l is the start-time (in frames) of p_l and $e \in \mathbb{N}$ is the end time of the last phone $p_{L^{\overline{v}}}$. For brevity we denote $s = s_1$.

A *spoken term detector* is a function $f : \mathcal{X}^* \times \mathcal{V}^* \rightarrow \mathbb{R} \times \mathbb{N} \times \mathbb{N}$, which takes as input the pair $(\overline{\mathbf{x}}, \overline{v})$ and returns a triplet: a real value expressing the confidence that the term $\overline{v}$ is uttered in $\overline{\mathbf{x}}$, the start time s and end time e of $\overline{v}$ in $\overline{\mathbf{x}}$. The confidence score output by f can be compared to a threshold $b \in \mathbb{R}$ to actually predict whether the term is uttered in $\overline{\mathbf{x}}$. In this work we ignore the start and end times and consider only the correctness of the prediction of existence or non-existence of the term in the input signal. For our purposes, therefore, we consider the detector to be a function $f : \mathcal{X}^* \times \mathcal{V}^* \rightarrow \mathbb{R}$ that returns a confidence only.

The detector f should be able to detect *any* term, including terms that might not have been seen in training. In what follows, we also assume that the input $\overline{\mathbf{x}}$ is an utterance short enough for any term of interest to occur at most *once*. This is not a restrictive assumption, since for longer signals, the detector may be applied in a sliding window of appropriate length on overlapping portions of the utterance.

3. DISCRIMINATIVE MODEL AND TRAINING

In this section, we describe a discriminative algorithm for learning a STD function from a training set. The goal of any discriminative algorithm is to find a function that maximizes a given measure of performance. In STD, performance is often measured in terms of the receiver operating characteristic (ROC) curve of the true-positive (detection) rate versus the false-positive rate. Each point on this curve represents an operating point of the system. The average performance over all operating points is the Area Under the ROC Curve (AUC), ranging from 0.5 (chance performance) to 1 (perfect detection). We propose an algorithm that aims at finding the STD function parameters so as to maximize the AUC on unseen data. The method presented here can be adapted to other evaluation functions, such as the *occurrence-weighted value* or the *actual term-weighted value (ATWV)* used in the 2006 STD NIST evaluation [1] or the Figure of Merit (FoM) [3].

The STD function $f_{\mathbf{w}}$ in this work is parameterized by a vector $\mathbf{w} \in \mathbb{R}^n$ of importance weights ("model parameters"):

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

where $\boldsymbol{\phi} \in \mathbb{R}^n$ is a vector function composed of a set of predefined feature maps $\{\phi_j\}_{j=1}^n$, where $\phi_j : \mathcal{X}^* \times \mathcal{V}^* \times \mathbb{N}^* \rightarrow \mathbb{R}$. That is, each feature map takes as input the acoustics $\overline{\mathbf{x}}$, the term $\overline{v}$, and a proposed sequence of phone start and end times $\overline{s}$, and returns a scalar which, intuitively, represents the confidence that the term occurred in the proposed time span. The maximization over $\overline{s}$ corresponds to finding the best scoring occurrence of the term $\overline{v}$ in the input utterance $\overline{\mathbf{x}}$, and this best score is the confidence returned by $f_{\mathbf{w}}$. The maximization defined by (1) is over an exponentially large number of time spans and alignments. Nevertheless, as in HMMs, if the feature maps $\boldsymbol{\phi}$ are decomposable, the maximization can be efficiently calculated via dynamic programming.

3.1. Feature maps

In this section, we describe two types of feature maps used that we use in our system. As stated before, we assume that the phone sequence $\overline{p}$ is obtained via the dictionary function $\overline{p} = \pi(\overline{v})$. The feature maps are constructed using a set of *feature functions* that are extracted from the acoustic feature vectors and can incorporate information from many diverse sources. For example, these may be posterior probabilities from classifier outputs, Gaussian likelihoods, or tandem-style feature representations [4]. In this work we use two vector-valued functions $\boldsymbol{\xi}_1 : \mathcal{X} \rightarrow \mathbb{R}^{r_1}$ and $\boldsymbol{\xi}_2 : \mathcal{X} \rightarrow \mathbb{R}^{r_2}$, which take as input an acoustic feature vector corresponding to a frame of speech $\mathbf{x} \in \mathcal{X}$ and output a vector in $\mathbb{R}^{r_1}$ and $\mathbb{R}^{r_2}$.

The first type of feature map models the confidence of the phone label hypothesized for each frame of the target term:

$$\boldsymbol{\phi}_{1,q} = \frac{1}{e - s + 1} \sum_{l=1}^{L^{\overline{v}}} \sum_{t=s_l}^{s_{l+1}-1} \boldsymbol{\xi}_1(\mathbf{x}_t) \delta[p_t = q] \tag{2}$$

where $q \in \mathcal{P}$ is a particular phone and p_t is the hypothesized phone at frame t given the term $\overline{v}$ and the hypothesized start and end times. Thus, we have a set of $|\mathcal{P}|$ feature maps, each of which is a vector-valued function of the same length as $\boldsymbol{\xi}_1$.

The second set of feature maps model the acoustics at phone transitions for each pair of phones $q, q' \in \mathcal{P}$:

$$\boldsymbol{\phi}_{2,q,q'} = \frac{1}{e - s + 1} \sum_{t=s}^{e-1} \boldsymbol{\xi}_2(x_t) \delta[p_t = q \wedge p_{t+1} = q'] \tag{3}$$

where $q, q' \in \mathcal{P}$. Thus, we have a total of $|\mathcal{P}|^2$ feature maps of the second type, for each pair of phones, each of which is a vector-valued function of the same length as $\boldsymbol{\xi}_2$. As described in Section 4, our $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are post-processed posteriors produced by multilayer perceptron phone and articulatory feature classifiers.

3.2. Large-margin training

Recall that our goal is to find the weight vector $\mathbf{w}$ that maximizes the expected AUC. Assume that each triplet $(\overline{v}, \overline{\mathbf{x}}^+, \overline{\mathbf{x}}^-)$ is drawn from a fixed but unknown distribution ρ, where $\overline{\mathbf{x}}^+$ and $\overline{\mathbf{x}}^-$ represent utterances where the term $\overline{v}$ is present and absent respectively. The goal can be written in the form of the *Wilcoxon-Mann-Whitney statistic* [5] as

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \ \mathbb{P}\Big[f_{\mathbf{w}}(\overline{\mathbf{x}}^+, \overline{v}) > f_{\mathbf{w}}(\overline{\mathbf{x}}^-, \overline{v})\Big] \tag{4}$$

$$= \arg\max_{\mathbf{w}} \ \mathbb{E}\Big[\delta[f_{\mathbf{w}}(\overline{\mathbf{x}}^+, \overline{v}) > f_{\mathbf{w}}(\overline{\mathbf{x}}^-, \overline{v})]\Big], \tag{5}$$

where the probability and expectation are taken with respect to $(\overline{v}, \overline{\mathbf{x}}^+, \overline{\mathbf{x}}^-) \sim \rho$. Since we do not know ρ, we use a training set $\mathcal{T}$ of m examples drawn from the same distribution, $\mathcal{T} = \{\overline{v}_i, \overline{\mathbf{x}}_i^+, \overline{\mathbf{x}}_i^-, s_i^+, e_i^+\}_{i=1}^m$, where each example is composed of a term $\overline{v}_i \in \mathcal{V}^*$, a positive utterance $\overline{\mathbf{x}}_i^+ \in \mathcal{X}_{\overline{v}_i}^+$ in which the term $\overline{v}_i$ is uttered, a negative utterance $\overline{\mathbf{x}}_i^- \in \mathcal{X}_{\overline{v}_i}^-$ in which the term $\overline{v}_i$ is not uttered, and the time span (first and last frames) (s_i^+, e_i^+) of the term $\overline{v}_i$ in $\overline{\mathbf{x}}_i^+$. We do not put any restriction on the terms and allow $\overline{v}_i = \overline{v}_j$ for some i and j.

Maximizing the AUC is equivalent to minimizing the expectation over $\delta[f_{\mathbf{w}}(\overline{\mathbf{x}}^+, \overline{v}) < f_{\mathbf{w}}(\overline{\mathbf{x}}^-, \overline{v})]$. This, in turn, is equivalent to minimizing the expectation over $\delta[f_{\mathbf{w}}(\overline{\mathbf{x}}^-, \overline{v}) - f_{\mathbf{w}}(\overline{\mathbf{x}}^+, \overline{v})]$. The structural hinge-loss is an upper bound to this term, and is defined as

$$\ell(\overline{v}, \overline{\mathbf{x}}^+, \overline{\mathbf{x}}^-, \mathbf{w}) = \Big[1 - f_{\mathbf{w}}(\overline{\mathbf{x}}^+, \overline{v}) + f_{\mathbf{w}}(\overline{\mathbf{x}}^-, \overline{v})\Big]_+ \quad (6)$$

where $[z]_+ = \max\{0, z\}$. Overall, the weight vector $\mathbf{w}$ is found by minimizing the regularized average structural hinge-loss over the training set:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \ \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \ell(\overline{v}_i, \overline{\mathbf{x}}_i^+, \overline{\mathbf{x}}_i^-, \mathbf{w}) \quad (7)$$

It is important to note that we assume that the location of the term $\overline{v}_i$ in the positive sentence $\overline{\mathbf{x}}_i^+$ is known, but that the start and end times of individual phones within the term are unknown. Thus, we express $f_{\mathbf{w}}(\overline{\mathbf{x}}_i^+, \overline{v}_i)$ as $\max_{\overline{s}^+} \mathbf{w} \cdot \boldsymbol{\phi}(\overline{\mathbf{x}}_i^+, \overline{v}_i, \overline{s})$, where the maximization is over the set of start-time sequences restricted to the term time-span given in the training set. On the other hand, we express $f_{\mathbf{w}}(\overline{\mathbf{x}}_i^-, \overline{v}_i)$ as $\max_{\overline{s}^-} \mathbf{w} \cdot \boldsymbol{\phi}(\overline{\mathbf{x}}_i^-, \overline{v}_i, \overline{s}^-)$, where no restriction is imposed on the possible start-time sequence.

The objective in (7) is a difference of convex functions and hence can be optimized by the convex-concave computational procedure (CCCP) [6]. Pseudocode of the training algorithm is given in Figure 2.

4. EXPERIMENTS

We present experiments on a subset of the Switchboard corpus [7], using varying training set sizes. We first select all sentences in sets 23–49 containing at least four words other than non-speech sounds. From this candidate set, we build four training corpora of increasing size containing 500, 1000, 2500, and 5000 sentences, such that each corpus is included in the next larger set. We construct a 40-keyword set for tuning and a 60-keyword set for final testing by selecting words from sets 20–22 that occur at least five times in Switchboard. For each keyword, we select 20 sentences containing the keyword (positive sentences) and 20 sentences not containing the keyword (negative sentences) to obtain corresponding development and test sets. We remove initial and final silences from all utterances in the train, development and test sets.[1]

[1]Details are available at `http://www.ttic.edu/keshet/Source_Code.html`

Input: training set $\mathcal{T} = \{\overline{v}_i, \overline{\mathbf{x}}_i^+, \overline{\mathbf{x}}_i^-, s_i^+, e_i^+\}_{i=1}^m$; parameter λ

Initialize: $\mathbf{w}_0 = \mathbf{0}$

For $t = 1, \ldots, T$

  For $i = 1, \ldots, m$

    Predict: $\overline{s}_i^+ = \arg\max_{\overline{s}^+ : s=s_i^+, e=e_i^+} \ \mathbf{w}_{t-1} \cdot \boldsymbol{\phi}(\overline{\mathbf{x}}_i^+, \overline{v}_i, \overline{s}^+)$

  Set: $\mathbf{u}_0 = \mathbf{w}_{t-1}$

  For $j = 1, \ldots, J$

    Pick example $(\overline{v}_i, \overline{\mathbf{x}}_i^+, \overline{\mathbf{x}}_i^-, s_i^+, e_i^+)$, $1 \le i \le m$

    Predict: $\overline{s}_i^- = \arg\max_{\overline{s}^-} \ \mathbf{u}_{j-1} \cdot \boldsymbol{\phi}(\overline{\mathbf{x}}_i^-, \overline{v}_i, \overline{s}^-)$

    Set: $\Delta\boldsymbol{\phi}_i = \boldsymbol{\phi}(\overline{\mathbf{x}}_i^+, \overline{v}_i, \overline{s}_i^+) - \boldsymbol{\phi}(\overline{\mathbf{x}}_i^-, \overline{v}_i, \overline{s}_i^-)$

    Set: $\alpha_i = \min\left\{\frac{1}{\lambda}, \ \frac{[1 - \mathbf{u}_{j-1} \cdot \Delta\boldsymbol{\phi}_i]_+}{\|\Delta\boldsymbol{\phi}_i\|^2}\right\}$

    Update: $\mathbf{u}_j = \mathbf{u}_{j-1} + \alpha_i \Delta\boldsymbol{\phi}_i$

  Update: $\mathbf{w}_t = \frac{1}{J}\sum_{j=1}^{J} \mathbf{u}_j$

Output: The last weight $\mathbf{w}_T$.

Fig. 2. CCCP algorithm to optimize (7).

For each sentence in a training corpus, we select each word v_i that contains at least 5 phonemes in its canonical pronunciation as a candidate term, and we select the corresponding utterance as an instance of a positive example $\overline{\mathbf{x}}_i^+$ for that term. We randomly select a sentence from the training corpus that does not contain the keyword as a negative example $\overline{\mathbf{x}}_i^-$. The set so selected serves as a training set for the discriminative spoken term detection systems.

We compute the functions $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_r]$ following the basic methodology outlined in [8]. We train four multilayer perceptrons (MLPs), three of which are frame classifiers of articulatory features: lip configuration (L, 8 labels), tongue configuration (T, 25 labels), and glottis-velum (G, 5 labels). The remaining MLP is a phonetic frame classifier. Unlike the work in [8], these MLPs are trained on all phonetically transcribed data from sets 23–49 of the Switchboard Transcription Project (STP) data [9] using the Quicknet toolkit [10]. We parameterize the acoustics using 12$^{\text{th}}$-order PLP coefficients with energy, deltas and double deltas to obtain a 39-dimensional input representation. The feature vectors for a given frame are concatenated with the four preceding and succeeding frames to obtain a 351-dimensional input representation to the MLPs. The MLPs are single hidden layer feed-forward nets, with a sigmoid activation function on hidden layer nodes and a softmax output function on the output layer nodes, and are trained to optimize a cross-entropy criterion. The number of hidden nodes is tuned on a held-out development set. Once the MLPs are trained, we compute log-posteriors for the data in the training, development and test sets for all four MLPs and project all of these log-posteriors down to their top 39 principal components using Principal Components Analysis (PCA) to obtain a *tandem feature* representation [4]. These features serve as the obser-

System	500	1000	2500	5000
HMM	0.810	0.827	0.842	0.855
SystemA	0.870	0.882	0.901	0.915
SystemB	0.874	0.901	0.914	0.926

Table 1. Test set average AUC for the baseline HMM-based system and the two discriminative systems.

vations modeled using a mixture of Gaussians in our baseline HMM systems and are also used in the feature functions $\boldsymbol{\xi}$ of the discriminative systems.

The proposed systems are evaluated against a baseline context-independent HMM recognition network consisting of a keyword model - formed by concatenating 3-state HMM phone models corresponding to the pronunciation of the term $\overline{p}^{\overline{v}}$ - and a garbage model consisting of all phone models in parallel. Given a test utterance, we compute the one-best Viterbi path through the network, which either passes through the keyword model (a detection) or passes solely through the garbabe model (a non-detection). The trade-off between the true positive and false positive rates is set by varying the keyword insertion probability. The baselines systems are trained using HTK [11]. The number of Gaussian components per mixture was tuned using the development set. The baseline HMM system trained on the 500 sentences employed 32 Gaussian components per mixture. All other systems are reported with 64 Gaussian components per mixture since in pilot experiments, adding additional Gaussian components did not result in significant performance improvements.

We evaluate two types of discriminative systems (SystemA and SystemB) that differ in their feature functions. In both systems, $\boldsymbol{\xi}_1(x)$ is the 39-dimensional tandem features concatenated with a bias term (a scalar constant). For the second set of feature functions, in SystemA $\boldsymbol{\xi}_2^A(x)$ consists of only a single constant bias term. In SystemB we include the phone transition-dependent functions, by setting $\boldsymbol{\xi}_2^B(x) = \boldsymbol{\xi}_1(x)$, which results in a 40-dimensional representation.

Table 1 shows results in terms of average AUC across all of the terms in the test set. Both SystemA and SystemB outperform the HMM baseline by large margins. Although the performance of both the HMM-based and discriminative systems improves with increasing training set size, the discriminative systems even at very low training set sizes performs comparably to the HMM baseline trained on much larger data sets. Note that the discriminative systems have much fewer parameters than the baseline HMM-based systems. The performance of both discriminative systems is significantly better than the baseline ($p \leq 0.001$) using the Wilcoxon signed-rank test across all training set sizes. Incorporating acoustic dependence on the transitions as in SystemB improves performance further over SystemA, significantly ($p \leq 0.001$) for all training set sizes other than 500.

5. DISCUSSION

We have presented a large-margin discriminative approach to spoken term detection based on optimizing the AUC, which significantly outperforms an HMM-based detection system on Switchboard conversational speech in limited data settings. Over a range of training set sizes from 500 to 5000 utterances, the large-margin system achieves roughly 0.06 higher AUC, with further improvement when additional transition features are added. The approach is general and can be applied in any spoken term detection experimental setup.

Future work includes incorporating additional feature functions, such as ones accounting for pronunciation variation. The Switchboard corpus data we have used is characterized by very high variability, so feature functions sensitive to correspondences between expected and observed pronunciations, including fine variations detected using our articulatory classifiers, may be helpful. Considering the low data requirements of our approach, another natural extension is to apply the approach to new languages or new acoustic conditions.

6. REFERENCES

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

[2] Joseph Keshet, David Grangier, and Samy Bengio, "Discriminative keyword spotting," *Speech Communication*, vol. 51, pp. 317–329, 2009.

[3] 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.

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

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

[6] A. Yuille and A. Rangarajan, "The convex-concave computational procedure (CCCP)," in *Proc. NIPS*, 2002.

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

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

[9] S. Greenberg, J. Hollenback, and D. Ellis, "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus," in *Proc. ICSLP*, 1996.

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

[11] 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.