



Detection-Based ASR in the Automatic Speech Attribute Transcription Project

Ilana Bromberg², Qiang Fu¹, Jun Hou³, Jinyu Li¹, Chengyuan Ma¹, Brett Matthews¹, Antonio Moreno-Daniel¹, Jeremy Morris², Sabato Marco Siniscalchi¹, Yu Tsao¹, Yu Wang²

School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA¹
 Department of Computer Science and Engineering, Ohio State University, Columbus, OH, USA²
 Center for Advanced Information Processing, Rutgers University, New Brunswick, NJ, USA³
 {qfu, jinyuli, cyma, brett, antonio, marco, yutsao}@ece.gatech.edu, junhou@caip.rutgers.edu,
 {bromberg, morrijer, wangyub}@cse.ohio-state.edu

Abstract

We present methods of detector design in the Automatic Speech Attribute Transcription project. This paper details the results of a student-led, cross-site collaboration between Georgia Institute of Technology, The Ohio State University and Rutgers University. The work reported in this paper describes and evaluates the detection-based ASR paradigm and discusses phonetic attribute classes, methods of detecting framewise phonetic attributes and methods of combining attribute detectors for ASR.

We use Multi-Layer Perceptrons, Hidden Markov Models and Support Vector Machines to compute confidence scores for several prescribed sets of phonetic attribute classes. We use Conditional Random Fields (CRFs) and knowledge-based rescoring of phone lattices to combine framewise detection scores for continuous phone recognition on the TIMIT database. With CRFs, we achieve a phone accuracy of 70.63%, outperforming the baseline and enhanced HMM systems, by incorporating all of the attribute detectors discussed in the paper.

Index Terms: Detection-based ASR

1. Introduction

The most commonly adopted approach to the task of automatic speech recognition (ASR) is to train acoustic models for a prescribed alphabet of short linguistic units, usually at the subword level, and to use dynamic programming techniques to find the best sequence of words for a given spoken utterance. While much of the success in the performance of ASR systems is directly attributable to this paradigm and its variants, the approach has a number of important limitations. Particularly, it is *knowledge-ignorant* in that the acoustic models are trained by collecting large corpora of transcribed speech, and fitting the best parameters of a distribution to the data. As a result, a wide body of expert knowledge in linguistics and acoustic phonetics is largely unused in modern ASR systems.

In this paper we discuss the first 3 years of work toward a new, *detection-based* paradigm for ASR, proposed to address some of the limitations of modern ASR systems and to narrow the significant gap between ASR and human speech recognition. Specifically, we present methods of detector design in the Automatic Speech Attribute Transcription (ASAT) project, where we have incorporated detectors of various attributes of the speech signal (sometimes referred to in the literature as *distinctive features* or *phonological features* or *acoustic-phonetic features*) into our approach to ASR.

Figure 1 illustrates the detection-based ASR paradigm developed over the course of the ASAT project. At the front end is a bank of detectors of useful and meaningful attributes of the speech signal. The outputs of these detectors, typically confidence scores for each attribute, are fused to infer higher-level evidences for the speech recognition task. Our selection

of speech attributes is taken directly from the area of linguistics. Detection-based ASR then represents an opportunity to effectively and methodically incorporate expert knowledge of linguistics and acoustic phonetics into speech recognition systems.

We highlight previous and current work in the ASAT project as well as the results of a cross-site collaboration (led by student researchers) between Georgia Institute of Technology, The Ohio State University and Rutgers University. We also embody in this paper an appeal to the research community at large for collaboration and input.

The paper is organized as follows: in Section 2 we describe our methods for designing detectors of speech attributes. In Section 3 we evaluate the performance of the detectors and the detection-based ASR systems we have developed. We also discuss our methods for fusing detectors of speech attributes, namely conditional random fields and knowledge-based lattice rescoring, and their results in the continuous phone recognition (CPR) task. Section 4 discusses ASAT as an open, collaborative platform for research in ASR and encourages input from researchers in a wide variety of fields. Finally, conclusions and future work are given in Section 6.

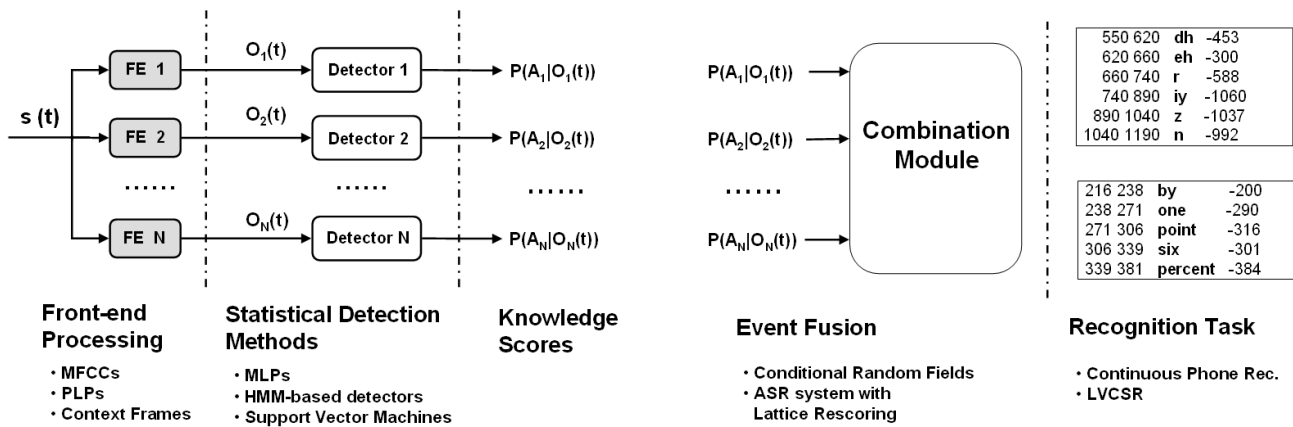
2. Designing speech attribute detectors

The front end of the ASAT detection-based ASR system is depicted in Figure 1 (a). The speech signal is first analyzed by a bank of detectors, each producing a confidence score or posterior probability pertaining to some acoustic-phonetic attribute. The design of these detectors, the optimization of their parameters and the selection of the set of attributes to detect are all critical design problems for the detection-based ASR paradigm. In this section we discuss our approaches to the design of speech attribute detectors.

The ASAT detection-based ASR system is designed to incorporate a wide variety of knowledge sources. Detectors of varying design methodologies and front-end processing techniques each have their own strengths and advantages and can be easily incorporated into our framework. A summary of our attribute detectors, developed independently at multiple remote sites, and the speech attribute classes we use, is given in Table 1. We use Multi-Layer Perceptrons (MLPs), Hidden Markov Models, Support Vector Machines (SVMs) and Time-Delay Neural Nets (TDNNs) to detect various acoustic-phonetic distinctive attributes of the speech signal. We also use MLPs to detect boundaries between phones and phonological features.

All of the speech attribute detectors discussed in this paper were trained on the TIMIT database.

10.21437/Interspeech.2007-510



(a) Front end.

Figure 1: ASAT Detection-Based ASR.

(b) Decoding.

2.1. MLP-based attribute detectors

2.1.1. MLP detectors for Sound Pattern of English classes

Using the Sound Pattern of English (SPE) features defined by Chomsky and Halle [2] as speech attributes, we built and optimized a set of Multi-Layer Perceptrons to detect each of the 14 binary-valued SPE features. The 61 TIMIT phonemes are mapped to the 14 SPE features, and the detection is done on each utterance frame by frame. We tested this architecture using both 2-layer and 3-layer MLPs using the Netlab and Matlab toolboxes. The input layer of the MLP has 13 nodes corresponding to 13 MFCC parameters in a single frame, and the output layer contains one node corresponding with one of the 14 SPE features[3].

2.1.2. Multiclass MLPs for Intl. Phonetic Assoc. (IPA) classes

Several multiclass MLPs, each with 1000 hidden nodes and between 3 and 9 output nodes, were used to detect 44 phonetic attributes as defined by the International Phonetic Association. The inputs are 13 perceptual linear predictive (PLP) features and their 1st- and 2nd-order time derivatives within a 9-frame window, including 4 frames each of preceding and following context. We trained 8 MLPs separately, each representing one phonological class from the IPA chart (sonority, voicing, etc) with several possible values. The **Voicing** class, for example, has labels: *Voiced*, *Voiceless* and *N/A*. These labels correspond to the three output nodes for the Voicing MLP. The *N/A* label is used to form an exhaustive class set when necessary.

Details of the eight MLPs used to detect the IPA attributes are given in the last row of Table 1. Collectively, the eight MLPs have 44 output nodes. We use each of these outputs as an individual attribute detector in the ASAT framework.

2.2. HMM-based attribute detectors

Conventional hypothesis testing is based on the Neyman-Pearson lemma which uses the likelihood ratio to accept or reject a proposed hypothesis. A generalized likelihood ratio is computed when a test observation \mathbf{O} is observed, and then compared against a decision threshold to decide which of two hypotheses is to be accepted. In order to conduct the test, one needs knowledge of two probabilistic models (for the null and alternative hypotheses), which are conventionally obtained through distribution estimation using pre-labeled data of sufficient amount. For the attribute detection problem, we model the null and alternative hypotheses with the well-known hidden Markov model.

We model each of the 17 phonetic attributes listed in Table 1 (last column, second row) with a *pair* of HMMs. Each target phonetic attribute and an “anti-target”, is modeled with a 3-state, 16-mixture HMM. A 2-class recognition is first performed on the speech signal, using just the target HMM, to obtain segments.

Both HMMs are then Viterbi-aligned to each segment. For an observation \mathbf{O} , the detector score is computed as the log-likelihood ratio $LLR(\mathbf{O}) = \log L(\mathbf{O}|\lambda_0) - \log L(\mathbf{O}|\lambda_1)$ where $L(\mathbf{O}|\lambda_0)$ and $L(\mathbf{O}|\lambda_1)$ are acoustic likelihoods of the target and anti-target models, respectively[4, 7].

2.3. SVM-based attribute detectors

Kernel Machines are an increasingly popular family of methods for pattern recognition. Among Kernel Machines, Support Vector Machines (SVM) are the most widely used, and have been applied to many pattern recognition problems, including speech recognition. In all Kernel Machine methods the input space \mathbb{R}^n is implicitly mapped to a high-dimensional feature space \mathbb{R}^{n^k} . According to Mercer’s condition, the inner product $\langle \phi(x), \phi(x_k) \rangle$, where $\phi(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^{n^k}$ is the nonlinear function mapping the input space to the feature space, can be computed through a kernel $K(x, x_k)$. Kernel machine methods can then use linear classification techniques involving inner products in a non-linear space, achieving impressive performance in many classification tasks.

We train an SVM classifier for each of the 17 phonetic attribute classes listed in Table 1. We use 13 MFCC coefficients and the 4 preceding and 4 following frames, giving a 9-frame window and a 117-dimension feature vector.

2.4. Phonetic boundary detection

Regions near phone boundaries and phonological feature boundaries may carry rich and important information for speech recognition. We attempted to extract boundary information and integrate this type of attribute into ASR systems as supportive information. Acoustic features (12th order PLP coefficients and their derivatives), and estimated probabilities of phones and phonological features were used as input features to our boundary detectors. For each of the 8 broad phonetic classes listed in Table 1 (last column, third row), a fully connected Multi-Layer Perceptron (MLP) with 4 output nodes was developed to detect transitions between the classes, resulting in 32 attributes for phonetic boundary detection. The 4 output nodes for each MLP classify a frame of speech as a Left Boundary (LB), Right Boundary (RB), Non-Left Boundary (NL) or Non-Right Boundary (NR)[9].

2.5. Other detection-based methods

Over the course of the ASAT project, we explored several other detection-based methods for ASR. We briefly summarize these in the following sections.

2.5.1. Whole-word detectors

Detector design at different levels is one of the key components within the framework of a detection-based ASR system. We have also incorporated acoustic-phonetic knowledge into the design of HMM-based detectors of whole words. We have used

Methods of Detection	Front-end Processing	Speech Attributes
MLP (SPE)	13 MFCCs 10 msec frames	SPE Classes: vocalic consonantal high back low anterior coronal round tense voice continuant nasal strident silence (14 attributes)
SVM	13 MFCCs 9 context frames 10 msec frames	coronal dent fricative glottal high labial low mid nasal roundminus roundplus silence stop velar voicedminus voicedplus vowel (17 attributes)
HMM-Based	13 MFCCs + Δ + $\Delta\Delta$ 10 msec frames	
Multi-class MLPs	13 PLPs + Δ + $\Delta\Delta$ 9 context frames 10 msec frames	Sonority: Obstruent Silence Sonorant Syllabic Vowel Voicing: NA Voiced Voiceless Manner: Approximant Flap Fricative NA Nasal NasalFlap Stop-Closure Stop Place: Alveolar Dental Glottal Labial Lateral NA Palatal Rhotic Velar Height: High Low-High Low Mid-High Mid NA Backness: Back Back-Front Central Front NA Roundness: NA NonRound NonRound-Round Round-NonRound Round Tenseness: Lax NA Tense (44 attributes)

Table 1: Summary of detectors, front-end processing methods and speech attributes

knowledge-based attributes in fine analysis and pruning for detected word segments. While word detection in this regard is similar to wordspotting, our word detectors work well on both content words and function words.

2.5.2. Time-delay neural nets for voiced-stop classification

Time-Delay Neural Networks (TDNN) have been shown to be effective in classifying dynamic sounds. For dynamic sounds such as the voiced stop consonants, their acoustic feature values change significantly over the duration of the sound. The TDNN can look for the desired feature changes by duplicated time-delayed weights, and thus encode the fine temporal structure within a segment. We built a set of TDNNs to classify the stop consonants of /b/, /d/, /g/ and /p/, /t/, /k/, which are among the most difficult consonant classes to classify accurately.

2.5.3. The role of finite state automata (FSAs)

The Finite State Automata (FSA) framework is an efficient and powerful representation of recognition networks, developed for ASR during the 1990s. FSAs are a convenient representation for the detection of multiple events, and can play an important role in detection-based ASR. *Transducers* in particular, are a type of FSA that allows assembly across layers of representation (acoustic models, phones, words). Similarly, sequences of detected attributes, at various levels of abstraction, could be represented as transducers.

3. Evaluating detection-based ASR

In this section we evaluate the stand-alone performance of our attribute detectors as well as the performance of our detection-based ASR systems. Specifically, we perform continuous phone recognition (CPR) and large-vocabulary continuous speech recognition (LVCSR) using conditional random fields (CRFs) and knowledge-based rescoring of speech lattices. All experiments are carried out on the TIMIT database.

3.1. Detector performance

A compelling advantage of the detection-based ASR paradigm and the use of bottom-up knowledge integration is that the stand-alone performance of low-level detectors of knowledge sources can be evaluated. In this section we briefly evaluate the performance of detectors of knowledge sources in the context of detection-based ASR.

The DET curve, much like the well-known Receiver Operating Characteristics (ROC) curve, plots the locus of a detector's accuracy over the complete range of threshold values. This kind of analysis can be used to examine the trade-off between the number of false alarms and false rejects a detector will produce, and was used extensively in the development of speech attribute de-

tectors in the ASAT project.

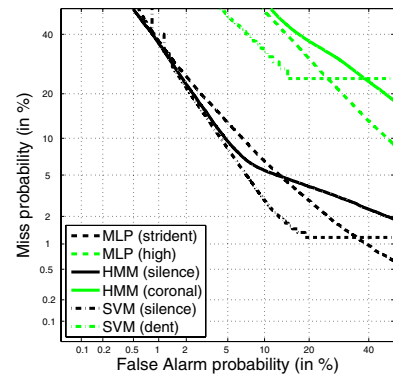


Figure 2: Selected Detector Error Trade-off (DET) curves for 2-class MLP, HMM and SVM detectors.

Selected plots of the Detector Error Trade-off (DET) curve are given in Figure 2. The plots represent the best and worst performing attributes for the HMM, 2-class MLP and SVM¹ detectors. HMMs and SVMs perform best for detecting *silence* while 2-class MLPs detect the *strident* attribute best. The best and worst equal error rates (EER), are given in Table 2. At the EER, the miss rate and false alarm rate are equal.

Methods of detection	max EER	min EER
2-class MLP	0.250 (high)	0.081 (strident)
HMM	0.286 (coronal)	0.069 (silence)
SVM	0.417 (mid)	0.060 (silence)

Table 2: Minimum and maximum Equal Error Rate (EER).

3.2. Continuous phone recognition

3.2.1. Conditional random fields

Conditional Random Fields (CRFs), are discriminative models for sequences that attempt to model the posterior probability of a label sequence conditioned on a set of input observations. A CRF defines the conditional probability $P(\mathbf{Y}|\mathbf{X})$ as: $P(\mathbf{Y}|\mathbf{X}) = \exp \sum_t (\sum_i \lambda_i f_i(\mathbf{Y}, \mathbf{X}, t)) / Z(\mathbf{X})$ where \mathbf{Y} is a sequence of labels, \mathbf{X} is a set of input observations, each function f is a feature function with an associated λ -weight, and the term $Z(\mathbf{X})$ is a normalizing term computed over all label sequences.

As in [6], we use CRFs to perform continuous phone recognition on the TIMIT speech database using only the attribute detectors discussed in the previous section as inputs. A non-linear

¹The worst performing SVM detector (mid) is not shown in Figure 2

Attribute Detectors	No. of Attrs.	Accuracy (%)	Correct (%)
Multi-class MLP (MC-MLP)	44	68.96	72.81
HMM	13	46.14	53.21
SVM	17	42.83	45.75
2-Class MLP (2C-MLP)	14	46.51	51.71
MC-MLP, HMM	44+13	68.56	73.95
MC-MLP, SVM	44+17	69.29	73.70
MC-MLP, 2C-MLP	44+14	69.15	74.26
MC-MLP, HMM, 2C-MLP	44+13+14	68.54	75.18
HMM, Phonetic Feature Boundaries (PFB)	13+32	51.50	57.47
MC-MLP, PFB	44+32	69.02	71.37
MC-MLP, SVM, PFB	44+17+32	69.26	71.34
MC-MLP, HMM, PFB	44+13+32	70.47	73.56
MC-MLP, HMM, 2C-MLP, PFB	44+13+14+32	70.63	74.48

Table 3: Continuous phone recognition experiments with CRFs on the TIMIT database

scaling was first applied to the HMM-based detector scores which are not strictly confined to a finite range. Table 3 gives results of continuous phone recognition experiments performed using CRFs with several configurations of speech attribute detectors as inputs. The results in Table 3 are sectioned into 3 groups. Experiments in the first group involve using just one of the sets of attribute detectors in Table 1. Among these, the best performance, a phone accuracy of 68.96%, is achieved with multi-class MLPs.

In the second and third groups in Table 3, two or more sets of attribute detectors are used as inputs to the CRF system. In the second group, all detectors are combined with multi-class MLPs, the best performing attribute detectors from the first group, and an accuracy of 69.29% is obtained in the best case. Finally, we incorporate phonetic boundary detectors in the last group in Table 3. The best phone accuracy result, 70.63%, is obtained when HMM-based detectors, multi-class MLPs, 2-class MLPs and phonetic feature boundaries are all incorporated, making use of 103 knowledge scores. The results of this first set of experiments are very encouraging since, in our detection-based framework, there is always room to incorporate more knowledge sources.

3.2.2. Knowledge-based lattice rescoring

While the ASAT detection-based ASR paradigm aims to perform automatic speech recognition using only detectors of speech attributes as inputs, it is instructive to use the same detectors to enhance the performance of state-of-the-art HMM-based speech recognition systems. We accomplish this for the continuous phone recognition task by rescoring phone lattices as in [7]. We use a 32-mixture, HMM with 3 active states per phone to perform continuous phone recognition, and then use output scores from speech attribute detectors to rescore lattices. Table 4 gives CPR results of the baseline system and after rescoring with HMM-based and multi-class MLP attribute detectors. The baseline performance of 61.16% improves to 64.59% with multi-class MLPs, which is comparable to [7] and falls slightly with HMM-based detectors. It is encouraging that the strictly detection-based ASR system in Section 3.2.1 outperforms a state-of-the-art HMM ASR system, even with these enhancements.

4. Collaborative Platform for Research

Central to the ASAT project is the goal of fostering an open platform for collaborative research. While we have developed a rich set of statistical methods and front-end processing techniques for

Attribute Detectors	No. of Attrs.	Acc (%)	Corr (%)
Baseline performance	---	59.48	63.02
HMM	13	59.17	62.79
Multi-class MLP	17	61.16	64.59

Table 4: Continuous phone recognition results with knowledge-based lattice rescoring on the TIMIT database.

attributes of the speech signal, our work represents only the first step toward building a community for detection-based ASR research.

With this work, we invite the collaborative efforts of the research community at large. As part of the ASAT project, we will develop online tools and standards which researchers can use to submit their detector designs and evaluate them in the context of our detection-based ASR system. The expert knowledge of researchers in ASR, as well as in linguistics, acoustics, signal analysis, statistical pattern recognition and other areas can then be easily incorporated into the ASAT project, providing a lasting framework for detection-based ASR research.

5. Conclusions and future work

We have presented some of the methods of speech attribute detector design developed during the first 3 years of the Automatic Speech Attribute Transcription project. We have applied MLPs, SVMs, and HMMs toward the detection of acoustic-phonetic attributes of speech taken from the linguistics research community. We have shown that these attributes can be combined to provide higher-level evidences useful for the speech recognition task. Specifically, we have performed continuous phone recognition using only knowledge-based scores as inputs. We have also integrated detectors of speech attributes into state-of-the-art HMM-based ASR systems.

Future work for the ASAT project involves building an open community for collaborative research. Researchers from a variety of areas are invited to use the tools for detection-based ASR and information exchange that we have developed and will make public. We believe that, through the detection-based ASR paradigm, input from experts in linguistics, signal analysis and other fields will be helpful in closing the performance gap between ASR and human speech recognition.

6. Acknowledgements

This work is supported by the National Science Foundation under NSF ITR grant IIS-04-27413.

7. References

- [1] Q. Fu and B.-H. Juang, Segment-Based Phonetic Class Detection Using Minimum Verification Error (MVE) Training.
- [2] N. Chomsky and M. Halle. *The Sound Pattern of English*. MIT press, 1991.
- [3] J. Hou, L. R. Rabiner, and S. Dusan. Automatic speech attribute transcription (asat) - the front end processor. In *Proc. ICASSP-2006*.
- [4] Jinyu Li, Yu Tsao, and Chin-Hui Lee. A study on knowledge source integration for candidate rescoring in automatic speech recognition. In *Proc. ICASSP-05*.
- [5] C. Ma, Y. Tsao, and C.-H. Lee. A study on detection based automatic speech recognition. In *Proc. InterSpeech-06*.
- [6] Jeremy Morris and Eric Fosler-Lussier. Combining phonetic attributes using conditional random fields. In *Proc. InterSpeech-06*.
- [7] Sabato Marco Siniscalchi, Jinyu Li, and Chin-Hui Lee. A study on lattice rescoring with knowledge scores for automatic speech recognition. In *Proc. InterSpeech-06*.
- [8] Yu Tsao, Jinyu Li, and Chin-Hui Lee. A study on separation between acoustic models and its applications. In *InterSpeech-05*.
- [9] Y. Wang and E. Fosler-Lussier. Integrating phonetic boundary discrimination explicitly into hmm systems. In *Proc. InterSpeech-06*.