AUTOMATIC SPEECH ATTRIBUTE TRANSCRIPTION (ASAT) – THE FRONT END PROCESSOR

Jun Hou, Lawrence Rabiner, and Sorin Dusan

CAIP Center, Rutgers University, Piscataway, NJ 08854, U.S.A.
{junhou, lrr, sdusan}@caip.rutgers.edu

ABSTRACT

In this paper we discuss the design and implementation of the ASAT front end processing system, whose goal is to convert the speech waveform into a range of measurements and parameters which are then combined to form probabilistic attributes. The ASAT front end processing module utilizes a range of spectral and temporal speech parameters as input to a set of neural network classifiers to create sets of attribute probability lattices, based on either single frames or blocks of frames (segments). We test this architecture by using the 14 Sound Patterns of English (SPE) features as speech attributes. Without balancing the training data, the detection accuracies of 4 of the SPE features are above 90%, 2 features obtain between 80% and 90% detection accuracy, and 8 features have detection accuracies below 80%. With a novel method of balancing the feature training data, the performance of the neural networks improved significantly, with 6 features having detection accuracies above 90% and the remaining 8 features with detection accuracy above 80%.

1. INTRODUCTION

Knowledge-based and statistics-based approaches are two directions in Automatic Speech Recognition (ASR) and both have evolved over time [12]. Hidden Markov Model (HMM) based speech recognition techniques [11] have achieved great success for controlled tasks. However, when we require improved (closer to human) accuracy and robustness, the HMM algorithms gradually fail. Hence a need has emerged to incorporate higher level linguistic information into ASR systems in order to further discriminate between speech classes or phonemes with high confusion rates.

The Automatic Speech Attribute Transcription (ASAT) project [8] has the long term goal of improving the performance of ASR systems by utilizing linguistically based speech attributes and speech events in an architecture that integrates knowledge sources, models, data, and tools, ultimately combining the results with state-of-the-art HMM systems.

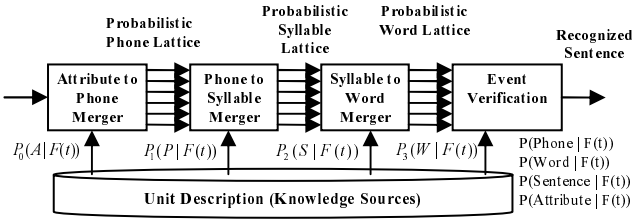

Fig. 1 Bottom-up knowledge integration

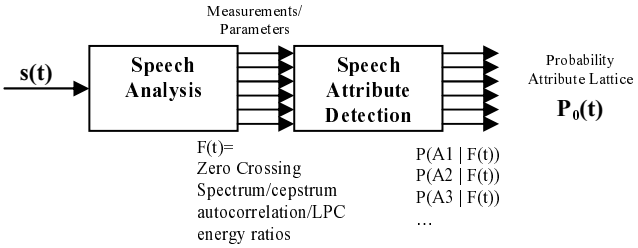

Fig. 2 Front end processing

In this paper we discuss the implementation of the ASAT front end processing system, whose goal is to estimate a set of attribute probability lattices $P_0(A/F(t))$, which can be combined with information from higher level knowledge sources (e.g., a word lexicon) to create a phone lattice $P_1(P/F(t))$, a syllable lattice $P_2(S/F(t))$ and a word lattice $P_3(W/F(t))$, which ultimately are used in a set of event verification modules to make the final recognition decision. Fig. 1 shows this process. Fig. 2 shows the general front end processing system. A speech parameter $F(t)$ is a direct measurement from the speech waveform, such as zero crossing rate or energy ratio. A speech attribute (also called speech evidence), A_i, is a piece of acoustic, phonetic or linguistic information that is estimated from the speech parameters. The attributes, e.g. voicing, nasality etc., distinguish the phonemes. An event, *e(t),* is a stochastic process corresponding to each attribute that is used to make the decision that either the attribute is present (+) or absent (–) at time *t*, as shown in Fig. 3. Such decisions can also be deferred to higher levels such as phones, syllables, words, and ultimately sentences, thereby mitigating the curse of

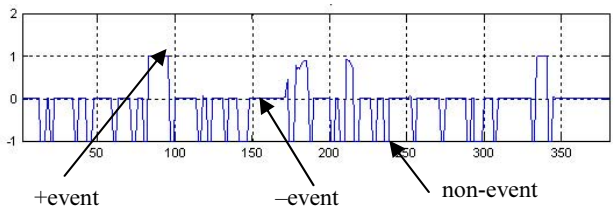

Fig. 3 Illustration of events

error propagation that has plagued linguistically based ASR systems over time.

There have been many research efforts that tried to utilize knowledge features in speech recognition. Morgan *et al.* summarize the state-of-the-art in this area in [9]. Landmark-based speech feature detection algorithms are described in [5]. The use of auditory models with specific measurements for stop and fricative consonants detection are described in [1, 2]. Almost all of these approaches are somewhat *ad hoc* in that they lack a general architecture for speech attribute detection.

2. IMPLEMENTATION OF THE ASAT FRONT END

The ASAT front end processing module utilizes a range of spectral and temporal speech parameters (both short-time and long-time measurements) as input to various sets of attribute classifiers (Bayesian Classifier, Multi-Layer Perceptrons (MLP), etc.) to create sets of attribute probability lattices, based on either single frames or blocks of frames (segments). Fig. 4 shows a block diagram of the front end framework. Using this framework we attempted to answer the following key questions, namely:

(1) ***What parameters to measure***—There are numerous temporal and spectral speech parameter sets (see Table 1). We had to choose parameter sets that would be most effective in estimating the speech attributes of interest.

(2) ***What signal processing algorithm should be used for each parameter***—Often there exist multiple signal processing algorithms for different speech parameters, e.g. to calculate formants or pitch period we could use cepstral or LPC methods. We had to make choices for each parameter set.

(3) ***What attributes to estimate***—Depending on the speech representation, different attribute sets are meaningful, including SPE binary attributes, linguistic attributes such as nasality, frication, etc.

(4) ***How to optimize attribute calculation from training***—Attribute events are usually obtained by some type of probability estimation process, e.g., Multi-Layer Perceptron (MLP) or Karhunen-Loeve (K-L) expansion.

We use as a training and testing set, data from the TIMIT database. A key issue with TIMIT data is that the phonetic labels are known to contain some inherent errors, both in fine placement, and often in phonetic identification. As such, part of the effort was spent in deciding how to handle the TIMIT miss-alignment of label issue.

3. SPEECH PARAMETERS

3.1. Temporal vs. spectral, short-time vs. long-time

Table 1 shows examples of the four speech parameter classes that were investigated.

Table 1. Speech parameter groups

	Short-time	Long-time
Temporal	voiced/unvoiced/silence Pitch Segmental SNR	VOT burst duration unvoiced duration syllable duration
Spectral	MFCC Spectral flatness Relative band energies	delta(MFCC) delta-delta(MFCC)

3.2. Frame based vs. segment based

Frames are flexible and convenient to implement, and they characterize static (short-time) properties of speech. Segments normally cover longer time spans (order of 10 frames or longer) and characterize speech dynamics. A segment generally contains a variable number of frames. In this paper all of our results are frame-based.

4. SPEECH ATTRIBUTE CALCULATION

The topology of speech attributes can be parallel, hierarchical or combined. The hierarchical architecture, [6], is most efficient but suffers from the problem of error propagation from higher levels of attributes to lower levels. The parallel topology (as used in ASAT) avoids this problem, but assumes that all speech attributes are independent whereas in reality they are not. In this initial experiment we use the parallel attribute organization. We tested this architecture using the set of 14 Sound Patterns of English (SPE) [4] distinctive features as speech attributes.

The attribute combination module can use the same architecture for all the attributes, or tailor the estimation method for each attribute. We chose to train a separate frame-based ANN for each of the 14 SPE features.

5. EXPERIMENTS

Most of our experiments used single frame parameters (nominally a set of 13 Mel-Frequency Cepstral Coefficients (MFCC)), but we also did experiments with other speech parameter sets. We used MLPs with both one and two hidden layers for our experiments. In all the following experiments, the window length was 32 ms and frame rate was 100 Hz.

We first tested the effect of boundary frames on the attribute estimates since the TIMIT phoneme alignment has known errors and we only want to use the most reliable data

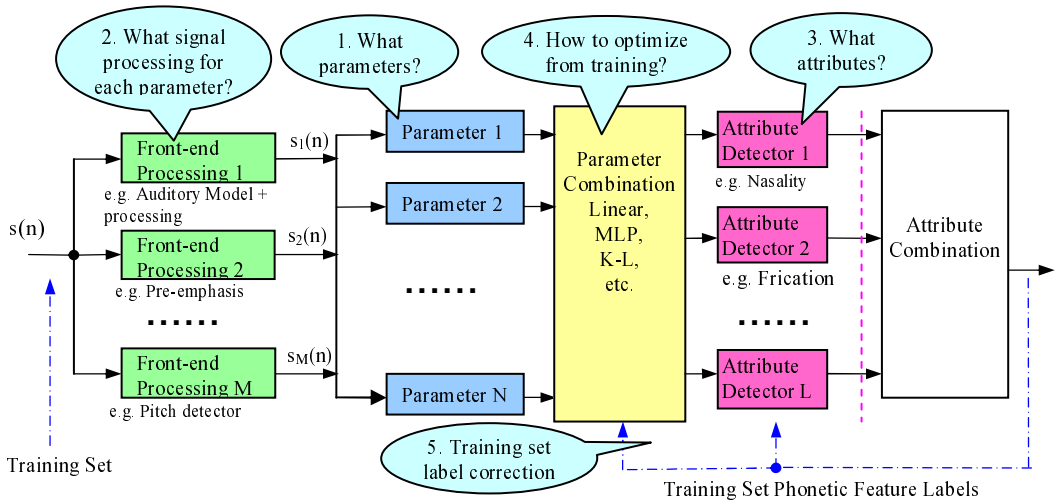

Fig. 4 Implementation of ASAT front end processing

for training. Using the Bayesian classifier described in [3] for voiced/unvoiced/silence frame classification and the entire TIMIT training set for training and test set for testing, we found that by restricting the training and test sets to a subset of the phonemes (we call it the "stable phoneme set" that includes vowels, diphthongs, semivowels, glides, fricatives, affricatives and silence, a total of 711357 frames for training and 262781 frames for testing), and by avoiding phone boundary frames we achieved classification accuracies of 99% for voiced frames, 87% for unvoiced frames, and 96% for silence/background frames. Using the model obtained above and testing on all phonemes in the test set (a total of 324229 frames, still omitting the unreliable phone boundary frames), the classification accuracy fell to 96% for voiced frames, 72% for unvoiced frames, and 93% for silence/background frames. Finally when all phonemes and all frames were used in testing (a total of 512536 frames), the classification error fell further to 93% for voiced frames, 60% for unvoiced frames, and 86% for silence/background frames.

We tested both 2-layer MLP using the Netlab toolbox [10] and 3-layer MLP using the Matlab neural network toolbox for classification of the 14 SPE features based on single frame 13 MFCCs. The 2-layer MLP consists of 100 nodes in the first layer and 1 node in the output layer. It performed well but there existed a convergence problem. Using the Matlab neural network toolbox, due to computer memory limits, we had to sample the frames in the training set. Due to the high correlation between adjacent frames, we choose 1 out of every four consecutive frames for training and testing. Thus, the training set size was 48,000 frames and the test set size was 33,020 frames. The phoneme boundary frames and the immediately adjacent frames were also discarded. For the 3-layer ANN we found that having 100 nodes at the first layer, 26 nodes for the second layer and 1 node for the output layer gave the best classification accuracy for the 14 SPE features.

Initially we trained the 3-layer MLP classifiers using randomly selected frames for each feature where there generally were far more occurrences of frames with the "– feature" present than frames with the "+ feature" present. We call this training set the "unbalanced" set. We classify the testing performance as "good" when the accuracy is above 90% for ***both*** + feature and – feature detection, as "acceptable" when both + and – feature detection rates are above 80% but at least one is below 90 %, or as "poor" when at least one of the feature detection rates is below 80%. The results for the unbalanced training set showed that for 4 of the 14 SPE features, the performance was "good", for 2 other features the performance was "acceptable", and for the remaining 8 features, the performance was "poor".

Since we are mostly interested in detecting the + features accurately and reliably, we devised a way to carefully balance the training set so that the number of training samples with the + features was comparable to the number of training samples with the – feature. Based on a balanced training set, the + feature detection performance significantly improved without seriously affecting the – feature detection accuracy. Our results showed that 6 MLPs achieved "good" detection performance (as compared to 4 for the unbalanced training set), and the remaining 8 features achieved "acceptable" performance (as compared to 2 for the unbalanced training set). For all the 14 features the average frame correctness for + feature detected as + is 90%, – detected as – is 90.5% and overall is 90.4% (as compared to 81.9%, 95.1% and 91.5% respectively for the unbalanced training set). Fig. 5 shows the comparison of detection performance on unbalanced and balanced training. On average there are 4 "+" features for a TIMIT phoneme. With a balanced training set, the "+" features can be detected more accurately and the area under an ROC curve

of a balanced training set is larger than that of any unbalanced training sets. King *et al.* [7] achieved similar results for the SPE feature detection, but they did not consider the importance of balancing the "+" and "–" features.

We also compared MFCC, PLP and RASTA-PLP speech parameters and found that MFCC coefficients gave the highest classification accuracies for 9 of the 14 SPE features, while PLP parameters gave the highest classification accuracy for 3 features, and finally RASTA-PLP gave the highest classification accuracy for the remaining 2 features.

6. CONCLUSION AND FUTURE WORK

During the first year of the ASAT project, we measured a range of spectral and temporal, short term and long term parameters, and included them in the ASAT parameter set. We extensively tested ANN's of different types and provided linguistic/distinctive feature labels with varying degrees of success. Training on balanced training sets showed significant improvements over standard ANN training methods which use randomly selected training data. Due to the TIMIT labeling errors, boundary frames were discarded for training purposes. We also found that different auditory models were of benefit to different speech features. In future research, we will investigate segment-based methods and compare their performance with that of frame-based methods. We also hope to find better speech measurement parameters, more meaningful attributes, and better parameter combination strategies to form attributes, in order to provide the next processing stages with more accurate detection probabilities.

7. ACKNOWLEDGEMENT

This work is supported under the NSF ITR grant, IIS-04-27413. The authors would like to acknowledge the contributions of our colleagues at the Georgia Institute of Technology (Chin-Hui Lee, Fred Juang, and Mark Clements), and at Ohio State University (Eric Fosler-Lussier and Keith Johnson).

8. REFERENCES

[1] A. M. A. Ali, J. Van der Spiegel, P. Mueller, "Acoustic-Phonetic Features for the Automatic Classification of Fricatives", *J. Acoust. Soc. Am*., Vol. 109, No. 5, pp.2217-2235, May 2001.

[2] A. M. A. Ali, J. Van der Spiegel, P. Mueller, "Acoustic-Phonetic Features for the Automatic Classification of Stop Consonants", *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 8, pp.833-841, Nov. 2001.

[3] B. S. Atal and L. R. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications in Speech Recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing,* Vol. ASSP-24, No.3, 1976.

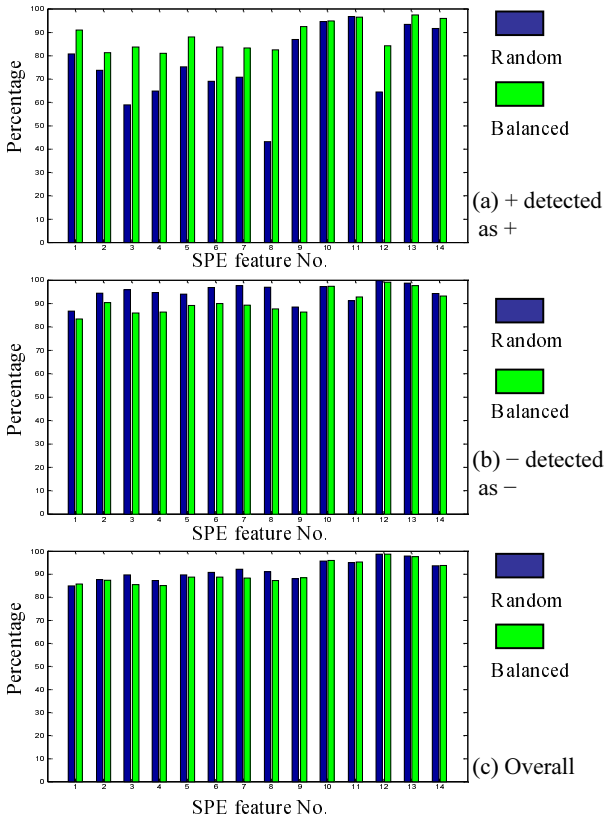

Fig. 5 Comparison of random training and balanced training for 14 SPE features. SPE No.1-vocalic, 2-consonantal, 3- high, 4-back, 5-low, 6-anterior, 7-coronal, 8-round, 9-tense, 10-voice, 11-continuant, 12-nasal, 13-strident, 14-silence.

[4] N. Chomsky and M. Halle, *The Sound Pattern of English*, MIT press, 1991.

[5] M. Hasegawa-Johnson, J. Baker, etc. "Landmark-based Speech Recognition: Report of the 2004 Johns Hopkins Summer Workshop", in Proc. *ICASSP 2005*, Philadelphia.

[6] A. Juneja, C. and Espy-Wilson, "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines", in the Proceedings of *International Joint Conference on Neural Networks*, Portland, Oregon, 2003.

[7] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks", *Computer Speech and Language* 14(4), pp. 333-353, 2000.

[8] C.-H. Lee, "From Decoding-Driven to Detection-Based Paradigms for Automatic Speech Recognition", in Proc. *ICSLP 2004*, Korea.

[9] N. Morgan, Q. Zhu, etc. "Pushing the Envelope – Aside", *IEEE Signal Processing Magazine*, pp.81-88, September 2005.

[10] I. T. Nabney, *NETLAB: Algorithms for Pattern Recognition*, Springer, 2001.

[11] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proc. *IEEE*, Vol. 77, No. 2, pp. 257-286, Feb. 1989.

[12] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.