INCORPORATING VOICE ONSET TIME TO IMPROVE LETTER RECOGNITION ACCURACIES

Partha Niyogi, Padma Ramesh

Bell Laboratories, Lucent Technologies
600 Mountain Avenue
Murray Hill, NJ 07974

ABSTRACT

We consider the possibility of incorporating distinctive features into a statistically based speech recognizer. We develop a two pass strategy for recognition with a standard HMM based first pass followed by a second pass that performs an alternative analysis to extract class-specific features. For the voiced/voiceless distinction on stops for an alphabet recognition task, we show that a linguistically motivated acoustic feature exists (the VOT), provides superior separability to standard spectral measures, and can be automatically extracted from the signal to reduce error rates by 48.7 % over state of the art HMM systems.

1. INTRODUCTION

There is little doubt that currently the most successful paradigm for speech recognition is a statistical approach typically using variants of an HMM framework [12]. While this approach has led to significant advances, some problems still remain. In this paper we investigate the possibility of using linguistically motivated features to correct some of the errors of current HMM based recognizers.

The notion of distinctive features [6] has long been regarded as a possible basis for automatic speech recognition. Unfortunately, few systems based on such principles have truly been implemented. Additionally, speech recognition research in this tradition has typically been conducted with hand-crafted rule-based approaches with relatively little statistical content to smooth over the inherent variability of the speech signal. At the same time, work in the mainstream statistical (primarily HMM based) approaches typically use a spectral sequence as features and ignore the possibility of linguistically motivated features. In our view, maximal benefits will emerge from a healthy union of statistical learning techniques with such feature systems. Our overall goal is to move towards such a feature based system. To demonstrate the feasibility of such a feature based approach, one will have to show that at least for one particular feature, a viable implementation exists. Specifically, one needs to ask the following questions: *What are the acoustic correlates of a particular distinctive feature? Do such acoustic correlates provide better separability than traditional spectral features (or transformations there-of like cepstra etc.)? Can such correlates be reliably extracted in an automatic speech recognition system?* This paper provides some answers to these questions on a limited task, i.e., alphabet recognition. As a starting point we examine the feature [voice] on stop consonants. As we shall see from an error analysis later, several of the errors in alphabet recognition occur due to a misclassification of this feature. To place our results in an appropriate context, it is worthwhile to emphasize some aspects of the work presented in this paper:

1. This paper should be viewed as a demonstration that at least for one particular case, i.e., the voiced/unvoiced distinction for stops in spoken letters, a linguistically motivated and perceptually real acoustical feature exists, can be automatically extracted and used for recognition leading to performance that is superior to state-of-the-art HMM systems. Very few such demonstrations exist. For example, the stop-recognition experiments conducted by Fanty and Cole (1990), Lamel (1988) Hasegawa-Johnson (1996) and Djezzar and Haton (1995) suggest that linguistic features provide reasonable performance but they have not been compared to state of the art HMM based systems (though Djezzar and Haton do show improvement of acoustic-phonetic features over cepstral features in a Neural Net setting). Furthermore, often the results have been presented on handsegmented speech. The same is true of the applications of feature based approaches to other sound classes ([11],[2]). (One notable exception, perhaps, are the promising results in Bitar and Espy-Wilson (1995)). At a time when many researchers are pessimistic about the future of acoustic-phonetic approaches, it is important to stress some of the positive results — the promising results on the voicing feature described in this paper suggests that it is worthwhile to investigate further the kinds of ideas discussed in Stevens (1995) and Zue (1985) where accounts of acoustical correlates of other phonetic distinctions have been presented.
2. We propose a two pass strategy for recognition. While the general idea of two pass strategies has been employed before in a number of different contexts, the details differ from system to system. In our case, we use a standard HMM based system as a first pass to obtain an initial tentative segmentation and classification of the speech signal. In the second pass, we employ a completely different analysis system that uses heterogeneous, acoustic-phonetic features to alter the segmentation and classification in a completely automatic manner. Since perceptual cues for recognition are presumably distributed in a non-uniform manner in the time-frequency plane, the second analysis system is crucial for improved and more biologically plausible recognition. The second pass recognizer is also statistically based: it builds probabilistic models on the new heterogeneous features.
3. We perform an analysis of the errors on a restricted alphabet task using a state of the art HMM system. We focus in particular on errors related to stops ("P","T","B","D","K") and their confusions. These are highly confusable sounds and require one to make fine phonetic distinctions that a human seems to make fairly easily while current recognition systems don't. We propose that the Voice Onset Time (VOT), an acoustically distinct and perceptually real

Authors in alphabetical order.

quantity, can be used as a criterion for discriminating the voiced from unvoiced stops (in pre-stressed, syllable initial position). This is a primarily temporal cue that is poorly modeled by current recognition systems. Most significantly, whenever the first pass HMM system classifies a segment as a stop, we invoke the second pass, automatically extract an estimate of the VOT and reclassify. Most statistical classifiers that depend on spectral distinguishability of the sound classes perform poorly on tasks such as that considered in this paper where the acoustic correlate seems to be primarily a temporal one.

4. It has been recognized in the past through the work of Lisker (1964), Klatt (1975) and others that the durational cue of Voice Onset Time provides good separation and is psychologically real. However, they have not addressed the issue of how such a measure can be automatically extracted from the signal; nor whether it provides superior separability to standard spectral measures. In the HMM tradition, VOT has not been considered as far as we know. In the acoustic-phonetic tradition, some recognition results exist using the VOT; mostly from hand-segmented speech; no account exists of how they compare with the performance of current HMM systems.

2. ERROR ANALYSES ON ALBHABET RECOGNITION

We considered a sub-problem of alphabet recognition on a database of spelled New Jersey town names spoken by 100 speakers (50 utterances each; 5000 utterances in all) and collected over the telephone. We concentrated in particular on voiced/voice-less minimal pair distinctions that need to be made for such alphabetical tasks.

2.1. Experience with Standard HMM Systems

We ran several variants of the standard HMM based recognizer (three state, left to right models) that have been trained on subword sequences with a front-end representation consisting of energy and cepstral coefficients and their first and second time derivatives. The basic cepstral representation was computed every 10 ms using a 30 ms window. The overall performance on the New Jersey townname alphabet set with a free grammar is a word (for the alphabet task, a word is the same as a letter) accuracy of 59.1%. The stops were the most confused words and accuracies were very low. Specifically, for the NJ townname alphabet task, some of the relevant letter accuracy rates are "T"(59.8%), "D"(54.8%),"B"(44.62%),"P"(66.67%).

These errors are seen to persist for most variants, e.g., when the subword models are changed from single phones to diphones with right context, or when duration models are used. It has been found from an analysis of confusion pairs, that the voicing dimension provides a high source of confusion for stops. Curiously, many more voiced stops are misclassified as unvoiced stops rather than the other way around. Table 1 shows the recognition scores for stops along with the number of confusions made along the voicing dimension for each stop. The results indicated are for the first pass system — a three state, left-to-right, subword based, HMM classifier. A large number of "A"'s were misrecognized as a stop ("K"); hence the inclusion of "A" results in the analysis.

2.2. Temporal versus Spectral Cues for Discrimination: the VOT

Why are the stop recognition scores so poor? For the task at hand, there are two primary factors. First, the traditional representation uses a 30 ms. analysis window moved every 10 ms. This is often too coarse to capture reliably the effects of a short duration, transient stop burst — typical voiced bursts are of the order of 5-10 ms. Second, even if the burst is captured reliably, a standard HMM system uses a purely spectral representation to classify sounds. In the case of the voiced/unvoiced distinction for letters, this is often unreliable. In this section, we provide some evidence that a temporal measure (specifically the Voice Onset Time) provides a greater degree of separation than purely spectral measures.

Alphabet	# Tokens	Corr.	Sub.	Confusion
T	2893	61.6 % 1782	30.5 % 884	1.5 % (T -> D) 42
D	1804	55.8 % 1006	34.2 % 617	16.6 % (D -> T) 304
P	1104	68.8 % 759	24.6 % 272	1.3 % (P -> B) 14
B	1163	44.8 % 521	46.4 % 540	14.9 % (B -> P) 173
K	1244	82.3 % 1024	11.7 % 145	0.2 % (K -> A) 2
A	4361	35.6 % 1554	36.4 % 1590	19.5 % (A -> K) 850

Table 1: Summary of relevant letter accuracies. The second column gives the total number of tokens for each letter in the database. The third column gives the number (and percentage) correctly recognized. The fourth column gives the number (and percentage) of that token that was substituted by some other in the recognition process. The remaining (column 2 - (column 3 + column 4)) provide the number of tokens deleted altogether. The number of tokens that were confused with voiced/unvoiced minimal pair is indicated in the fifth column. Note that the overall accuracy scores are computed after taking into account the number of incorrect insertions (not provided in this table) for each letter.

The canonical acoustic structure of a spoken letter containing a stop (i.e., "P", "T","B","D","K") consists of a closure, a burst release with frication and aspiration and the following vowel (E or A) as the case may be. The difference in time between the onset of the burst and the onset of voicing associated with the vowel is denoted as the Voice Onset Time (VOT). In syllable initial pre-stressed positions the VOT for unvoiced sounds is typically longer than that for voiced sounds. Various psychophysical studies have been conducted (notably by Lisker, 1975) where stimuli with varying VOTs were presented to subjects and their classification responses were measured. The VOT was found to be an important determinant in deciding phonetic class for stops.

Figure 1 shows the distribution of VOT values for "T" and "D" in syllable initial pre-stressed position. This has been obtained from data collected from syllable-initial stops excised from fluent speech using an inhouse database of 2000 phonetically balanced sentences spoken by a single male speaker. Notice the clear separation of the data with the unvoiced stop having a higher VOT in general than the voiced stop. This is consistent with previous literature and similar results exist for other stops as well; they have not been provided here for lack of space.

How well do "T" and "D" separate in the spectral domain? This is a trickier question to answer since it is difficult to compare across different distance metrics defined on different spectral

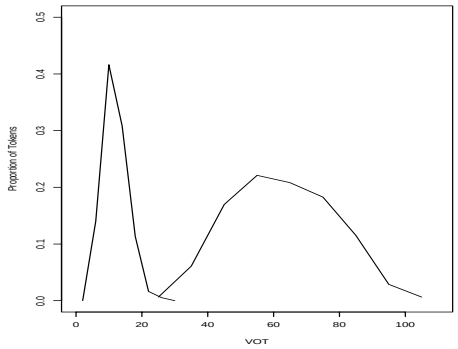

Figure 1: Distributions of VOTs for "T" (right) and "D" (left).

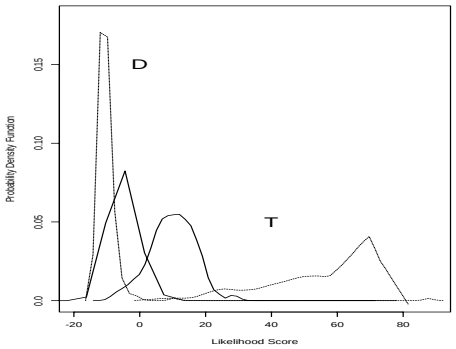

Figure 2: Separability of "T" from "D" using probability models constructed from spectral (solid) and VOT (dotted) measures. Notice the superior separability of VOT (indicated by the curves on the extreme left and right).

spaces of different dimensionalites. One way to get around this is by constructing probability models in the different feature spaces and using the following likelihood ratio discriminability measure:

For any x, define

$$d(x) = \log \frac{P(x|\Lambda_t)}{P(x|\Lambda_d)}$$

where $P(x|\Lambda_t)$ is the probability of an arbitrary point x in the feature space, given the model for "T" (Λ_t) constructed in that feature space (likewise for $P(x|\Lambda_d)$). Clearly $d(x)$ is large for points more likely to be generated by the model for "T" (likewise, small for "D"). By estimating the distributions of $d(x)$ for "T" and "D" tokens collected as before, we can characterize the separability for arbitrary feature spaces. Shown in fig. 2 are the distributions of $d(x)$ for probability models constructed in spectral space as well as probability models constructed in VOT space. The spectral representation consisted of filter bank outputs (logarithmically spaced). A principal components rotation was performed for orthogonalization and dimensionality reduction and Gaussian probability models were then constructed. In contrast, simple univariate Gaussian probability models were constructed in VOT space. Notice the significantly superior separability of the models developed using the VOT as a criterion. Similar results have been obtained with other kinds of spectral representations and have not been provided here for lack of space. This suggests that the VOT is a better candidate for an acoustic-phonetic feature that triggers this particular phonetic distinction. Psychophysical results of Lisker and others provide further credibility to this point of view.

3. RECOGNITION SETUP

Having demonstrated that the VOT provides better separability than usual spectral models, the important question remains: can one reliably extract it from the signal in an automatic manner and use it for superior recognition performance? We describe below one possible way in which this can be done.

3.1. A Two Pass Strategy

We have developed a two pass framework for recognition — the system diagram is shown in fig. 3. A standard HMM state-of-the art recognizer provides an initial recognition that is further refined using alternate features and classifiers. The second pass features and classifiers are appropriately tuned to specific sound classes and

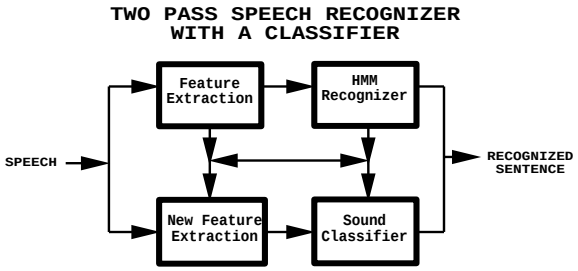

Figure 3: System outline for a novel two pass strategy incorporating alternative distinctive features in recognition. The upper half of the system corresponds to a standard HMM based recognizer that provides an initial segmentation and classification. The bottom half of the system diagram corresponds to the second pass that utilizes the output of the HMM as an initial guide and performs alternative feature processing that are tuned for confusible sound categories.

aim to reduce the errors made by the HMM. Most significantly, the second pass strategy allows for class-specific processing of temporal and spectral information in a more flexible manner.

As a first step, we have implemented such a strategy for letter classification with a second pass correcting only the confusions shown in table 1 . Due to the asymmetry in the confusion pair statistics, we targeted only those segments that were classified as an unvoiced stop by the standard classifier. Thus segments classified as "T","P" or "K" were reanalyzed in an attempt to locate the burst and voicing more precisely. The second pass took the speech segment that was classified as "T","P", or"K", and conducted a finer search to obtain a VOT estimate. A 1 ms. analysis was performed since considerable temporal precision is required for transient segments such as stops. An energy differential operator is used to locate the burst with a 1ms analysis. A pitch tracking algorithm using a normalized cross-correlation function with dynamic programming as in Talkin (1995) was used to locate the onset of voicing at a 10 ms rate. Thereby an estimate of the VOT was automatically computed from the signal. This VOT estimate was then used to reclassify the segment into the appropriate voiced/unvoiced category.

Alphabet	# Tokens	Corr.	Sub.	Confusion
T	2893	59.7 % 1726	32.5 % 940	3.3 % (T -> D) 98
D	1804	64.5 % 1164	25.4 % 459	8.1 % (D -> T) 146
P	1104	65.4 % 722	28.0 % 309	4.6 % (P -> B) 51
B	1163	57.2 % 665	34.0 % 396	2.5 % (B -> P) 29
K	1244	80.1 % 996	13.9 % 173	2.4 % (K -> A) 30
A	4361	46.9 % 2047	25.2 % 1097	8.2 % (A -> K) 357

Table 2: Summary of relevant letter accuracies after correction by the automatic second pass. As before, the second column gives the total number of tokens for each letter in the database, the third and fourth give the number of correct classifications and substitutions respectively and the final column gives the number of confusions made.

3.2. Results on Stops

As described earlier, the estimate of VOT was used to reclassify the stops into the corresponding voiced/unvoiced category. Different thresholds were picked depending upon the classified place of articulation of the stop. Note that by reclassifying in this manner, we now potentially misclassify some other previously correctly classified stops. We might also change some other confusions (e.g. Z $\rightarrow$ T might now be classified as Z $\rightarrow$ D). However, these new confusions do not affect the overall classification accuracy for the other sounds, it just affects the distributions of their errors. Shown in table 2 are the new confusion accuracies for the six letters under consideration here. Notice how (T -> D) has increased while (D -> T) has decreased.

In this experiment, we have targeted only the confusions caused by the voicing dimension for stops. Shown in table 3 are the six relevant confusions that our system handles at the moment. The three columns indicate the results for the first pass, the second pass with just one voiced/unvoiced classifier for all stops (using a VOT threshold of 40 ms. to classify) and a second pass with three classifiers with different thresholds depending upon the place of articulation of the relevant stop (best performance). Thus we see that the overall T/D confusions are reduced by 29%, the P/B confusions are reduced by 57%, and the K/A confusions are reduced by 54%. Thus for the six letters that we considered here, this amounts to an overall reduction of the error rate by 48.7%.

4. CONCLUSIONS

This is a first step towards incorporating distinctive features as an error correcting device to discriminate between confusable pairs in a statistical recognizer. From an examination of the voicing feature for stops, we conclude that the VOT, a temporal cue for discriminating between voiced and unvoiced stops in syllable initial positions provides superior separability to spectral cues. Furthermore, it can be extracted automatically from the signal and improves current recognition scores significantly. Future directions on the voicing feature include better ways of extracting the correlates of voicing in different contexts and testing robustness to noise. More significantly, we hope to enlarge the feature set to include a greater number of classes and the details of our second-pass classifier have to be developed further in this context.

Confusion	First Pass	Second P. (One)	Second P. (Three)
T -> D	42	81	98
D -> T	304	191	146
P -> B	14	76	51
B -> P	173	58	29
K -> A	2	18	30
A -> K	850	535	357

Table 3: Number of confusions along the voicing dimension using the first pass system; the second pass (with fixed VOT threshold set at 40 ms) and a second pass with place of articulation-dependent VOT threshold.

5. REFERENCES

[1] Bitar, N. and Espy-Wilson, C. "Knowledge-based parameters for HMM Speech Recognition," ICASSP, 1995.

[2] Espy-Wilson, C. "A Feature Based Approach to Speech Recognition", JASA, Vol.96,pp.65-72, 1994.

[3] L. Djezzar and J.P. Haton "Exploiting Acoustic-Phonetic Knowledge and Neural Networks for Stop Recognition," Eurospeech, 1995.

[4] Fanty, M. and Cole, R. "Speaker-Independent English Alphabet Recognition: Experiments with the E-set," ICSLP,1990.

[5] Hasegawa-Johnson, M. *Formant and Burst Spectral Measurements....*, Ph.D. Thesis, MIT, 1996.

[6] R. Jakobson, M. Halle, G. Fant, *Preliminaries to Speech Analysis*, 1952.

[7] D. H. Klatt, "Voice onset time, frication and aspiration in word-initial consonant clusters," *Journal of Speech and Hearing Res.*, vol. 18,pp.686-706, 1975.

[8] Lamel, L. F., *Formalizing Knowledge used in Spectrogram Reading....*, Ph.D. Thesis, MIT, 1988.

[9] Lisker,L and Abramson, A.S.,"A cross-language study of voicing in initial stops: acoustical measurements."*Word*, Vol. 20, pp. 384. 1964.

[10] Lisker, Leigh "Is it VOT or a first-formant transition detector," JASA, , 1975, pp. 1547-1551.

[11] H.M. Meng, V. W. Zue, and H. C. Leung,"Signal Representation, Attribute Extraction and the Use of Distinctive Features for Phonetic Classification," Fourth DARPA Speech and Natural Language Workshop, Feb., 1991, Pacific Grove, CA.

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

[13] Stevens, K.N.,"Applying Phonetic Knowledge to Lexical Access," *Proc. 4th European Conference on Speech Communication and Technology*, Vol. 1, pp. 3-11, Madrid, Spain, 1995.

[14] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal (ed)., Elsevier, 1995.

[15] Zue, V.W., "The Use of Speech Knowledge in Automatic Speech Recognition," *Proc. IEEE*, vol. 73, no. 11, 1985.