The Voicing Feature for Stop Consonants: Acoustic Phonetic Analyses and Automatic Speech Recognition Experiments

Padma Ramesh and Partha Niyogi

Bell Labs – Lucent Technologies
Murray Hill, NJ, USA.

ABSTRACT

We examine the distinctive feature [voice] that separates the voiced from the unvoiced sounds for the case of stop consonants. We conduct acoustic-phonetic analyses on a large database and demonstrate the superior separability using a temporal measure (*voice onset time*; VOT) rather than spectral measures. We describe several algorithms to estimate the VOT automatically from continuous speech and compare them on a speech recognition problem to reduce error rates by as much as 53 % over a baseline HMM based system.

1. INTRODUCTION

In this paper, we describe steps towards a distinctive feature based approach to speech recognition. In order to progress towards this goal, we need to better understand (1) the acoustic cues for such distinctive features (2) the reliability and separability of such cues particularly in comparison with the traditional acoustic representations such as cepstra (3) the mechanisms by which these cues can be automatically extracted and incorporated in an automatic speech recognition (ASR) system. Here, we resolve these issues for the case of the voicing feature for stop consonants. We focus on this feature partly because current ASR systems typically have greater error rates for these sounds.

Although the voicing feature for stop consonants manifests itself in complicated and context dependent ways, it has been observed in the phonetics literature [1] that an important cue consists of the voice onset time (VOT) that separates voiced stops from unvoiced stops, particularly in syllable-initial position. In recent work [6], we have incorporated VOT in an ASR system based on HMMs providing significant reduction in error rate along the voicing dimension. Here we expand on the theme of VOT and its use in speech recognition in a number of different ways.

While it is known that the VOT provides good separation between voiced and unvoiced stops, it is not known whether it provides better separation than the features traditionally used in ASR such as LPC-spectra or cepstra. Here we extend acoustic-phonetic analyses to multiple speakers on the TIMIT database demonstrating that the VOT is indeed much better than spectral separation, i.e., the cue to the voiced/unvoiced distinction is primarily a temporal rather than a spectral one.

In this paper, we also consider several alternative algorithms to estimate the VOT *automatically*. We describe the several versions of the VOT estimates, examine the accuracy of such estimates on unsegmented speech, and compare their performance on an ASR task. The best VOT estimation algorithm can reduce the error rate for the voiced/unvoiced distinction by as much as 53 % over a traditional HMM based system.

	#	P	T	K	B	D	A	E
P	1104	759	95	31	14	10	1	1
T	2893	198	1782	42	0	42	4	4
K	1244	14	71	1024	0	1	2	0
B	1163	173	76	8	521	69	1	13
D	1804	72	304	13	37	1006	0	0
A	4361	44	128	850	12	12	1554	48
E	5548	236	975	52	35	231	24	968

Table 1: Confusion Matrix with (i, j) element containing the number of $i \rightarrow j$ confusions. The total number of letters in the database is provided in column 1.

2. STOP CONSONANTS IN ALPHABET RECOGNITION

To ground our investigations in a concrete ASR problem, we consider a continuously spoken alphabet recognition task. This consists of recognizing spelled letters of New Jersey townnames continuously spoken by a hundred different speakers (50 utterances each; 5000 utterances in all) and collected over a telephone channel.

Our baseline system is an HMM based recognizer using 41 context independent phones. All phones were modeled by a three state, left to right HMM, except for silence which was modeled by a single state HMM. Models were trained using discriminative minimum classification error training on an inhouse general phrase and personal names database collected over the telephone [5]. Features used are energy and twelve cepstral coefficients (along with delta, and delta-delta values for these) computed every 10 ms using a 30 ms window. We focus in particular on letters containing stop consonants. Shown in table 1 is a subset of the overall confusion matrix for all alphabets that includes only the letters containing the stops and their confusions along the voicing dimension. We have included "E" and "A" here since confusion between stops and these letters is very high and are related partially to poor detection of a burst by our baseline system.

3. TEMPORAL VERSUS SPECTRAL CUES FOR STOPS

In this section, we provide evidence that a temporal measure like the VOT provides better separation along the voicing dimension for stops than typical spectral measures that are used in standard HMM based speech recognition systems.

3.1. Temporal Separability

To begin, let us consider a single speaker. Approximately two hundred "t"s and "d"s in syllable initial, pre-stressed position

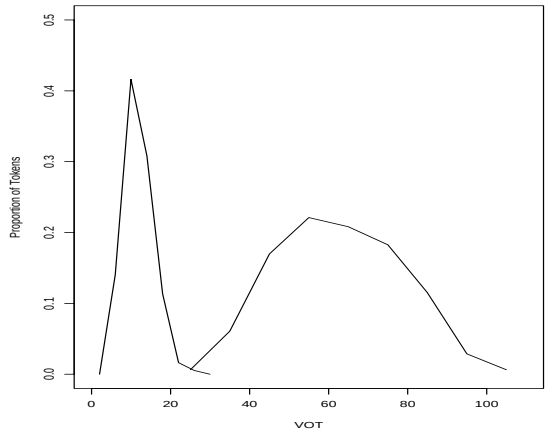

Figure 1: Distributions of VOTs for "t" (solid) and "d" (dotted); obtained from a single male speaker in several vocalic contexts.

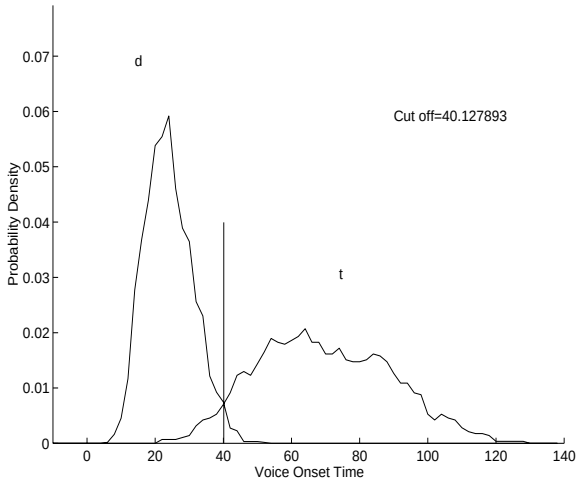

Figure 2: Distributions of VOTs for "t" and "d" in syllable-initial pre-stressed position extracted from the TIMIT database.

were extracted by hand from an inhouse database consisting of 2000 phonetically balanced sentences produced by a single male speaker (the KBB database). Shown in fig. 1 are the distributions of these VOTs and note the almost perfect separation found between "t"s and "d"s for this speaker.

Similar results exist for other minimal pairs as well ("p"/"b" and "k"/"g"). Do these results generalize to the multiple speaker case? Shown in fig. 2 are the distributions of VOTs for stops in syllable initial position extracted from the 630 speakers of the TIMIT database. By performing a textual analysis on the various sentences in the TIMIT database, only those stops that occured in syllable-initial, prestressed position were considered. This is a significantly larger number of speakers than has been considered before in similar acoustic-phonetic studies.

Again, similar results exist for the other stop minimal pairs and we do not report the figures here for lack of space. Our results here are consistent with previous observations [8, 4, 1] — (1) VOT values are larger for unvoiced stops than voiced; (2) VOT values vary according to place of articulation with labials having the smallest and velars having the largest; (3) separability is greatest and most reliable when the stops occur before stressed vowels and is less in other contexts.

3.2. Spectral Separability

The previous section merely reconfirms well known results in the acoustic-phonetic community that demonstrate large separability using VOT as an acoustic cue. Here we address an important question that has been inadequately treated in the past — how do the same sounds separate in the spectral domain and is there any competitive advantage to using temporal over spectral measures? This is a trickier question to answer since it is difficult to compare across different distance metrics defined on different spectral spaces of different dimensionalites. One way to get around this is by constructing probability models in the different feature spaces and using a likelihood ratio discriminability measure for voiced/unvoiced pairs as follows: For any x, define

$$d(x) = \log(\frac{P(x|\Lambda_u)}{P(x|\Lambda_v))}$$

where $P(x|\Lambda_u)$ is the likelihood of an arbitrary point x in the feature space, given the probability model for a particular unvoiced stop (like "p","t", or "k") constructed in that feature space (likewise $P(x|\Lambda_v)$ is the model for the voiced counter part, i.e., "b","d"and "g"). Clearly, for example, $d(x)$ is large for points more likely to be generated by the model for "t" (likewise, small for "d"). Thus, $d(x)$ is now a dimensionless quantity whose distributions characterize separability for arbitrary features x. Shown in fig. 3 are the distributions of $d(x)$ for "t" and "d" tokens extracted from the KBB database. One set of curves is obtained by constructing models in the VOT space; the other set of curves is obtained by constructing models in a spectral space. The spectral representation consisted of filter bank outputs (logarithmically spaced) computed every millisecond. A principal components rotation was performed for orthogonalization and dimensionality reduction and Gaussian probability models (diagonal covariance matrix) were then constructed in the rotated space. Notice the significantly superior separability of the models developed using the VOT as a criterion.

Again, we need to see if this fact generalizes to the multispeaker case. Fig. 4 shows the separability between "t" and "d" using data collected from the TIMIT database (velar and labial show similar characteristics but have not been included for lack of space). The VOT separability curves were obtained by constructing probability models using the data shown in the previous section. The spectral representation for the TIMIT speakers consisted of the first twelve cepstral coefficients obtained from 30 ms. windows moved at a 10 ms. rate in the burst region of the respective stops. This is the traditional basis for the acoustic representation for speech in many speech recognizers including versions that have existed at Bell Laboratories. Once again, the same pattern unfolds. The separability in the VOT space is considerably larger than the separability in the cepstral space. One begins to notice now the considerable overlap between the voiced and unvoiced stops in the cepstral space. This results in high error rates along the voicing dimension for most tasks including the spoken letter recognition

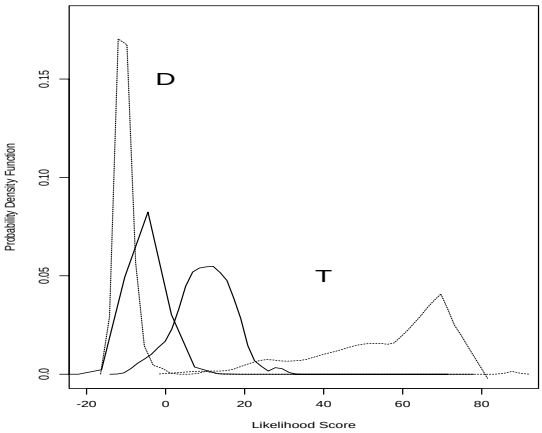

Figure 3: Separability of "t" from "d" using probability models constructed from spectral (solid) and VOT (dotted) measures. (Single Speaker; KBB).

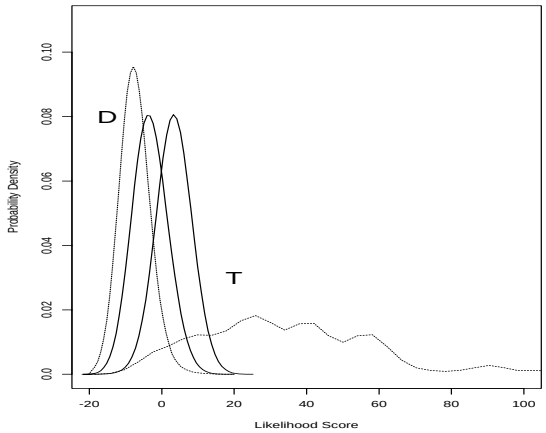

Figure 4: Separability of "t" from "d" using probability models constructed from spectral (solid) and VOT (dotted) measures (Multiple Speaker; TIMIT).

task considered in this paper.

4. AUTOMATIC VOT ESTIMATION

How do we automatically extract estimates of the VOT? We describe below 3 different methods of burst detection that were combined with a pitch tracker to yield an automatic estimate of the VOT. As we have earlier described in [6], we operate in a two pass manner. In a first pass using our baseline HMM system, we obtain a tentative segmentation of the signal yielding candidate segments where stops are postulated. In each such candidate segment, we now perform a detailed second pass analysis to locate estimates of two times: t_b (time at which closure-burst transition occurs) and t_v (time at which voicing comes on). Then $t_v - t_b$ yields an estimate of VOT.

Estimates of t_v are obtained by using a cross-correlation based pitch tracker with dynamic programming as in [7]. We describe below 3 different ways of obtaining an estimate of t_b. In what follows, we let $s(t)$ be the speech samples, and correspondingly let $E_t(n)$ be the total power (in dB) computed every 1 ms and $E_h(n)$ be the total power above 3 kHz (in dB) computed every 1 ms respectively. Therefore, t_b is in units of milliseconds from the start of the segment.

Algorithm 1: Optimized Differential Energy Operator. Here $t_b = \arg\max_n E_t(n) * h(n)$. The linear filter $h(n)$ satisfies $h(n) = 0 \forall n > 10$. The coefficients of the filter are estimated from training data using LMS training.

Algorithm 2: Optimized Linear Operator (Total Energy and H.F. Energy). Here $t_b = \arg\max_n (E_(n) * h_1)^2 + (E_h(n) * h_2)^2$. Both h_1 and h_2 are filters satisfying $h_i(n) = 0 \forall n > 10$. Their parameters are jointly estimated from training data using LMS training.

Algorithm 3: State Dependent Energy based Detector. Define state-variables $s(n)$ and $f(n)$ taking values in $\{0, 1\}$. Initialize $f(n) = 1$ and $s(n) = 0$ for all n.

for each n:

if $f(n)s(n) = 0 \& E_t(n) > th1 \quad s(k) = 1 \forall k \geq n;$
if $f(n)s(n) = 1 \& E_t(n) < th2 \quad s(k) = 0 \forall k;$
if $f(n)s(n) = 1 \& E_t(n) > th3 \quad f(k) = 0 \forall k > n;$

end

Finally, we let $t_b = \arg\max_n s(n) - s(n-1)$. In this algorithm the thresholds $th1, th2, th3$ are chosen from training data.

In the next section, we compare recognition results obtained by utilizing the above VOT estimation algorithms as a second pass to correct errors along the voicing dimension for letters. As we shall see then, *Algorithm 3* provides the best results in our recognition experiments. Of course, in the use of the VOT estimation algorithms in the two pass mode of section 5, it often happens that the postulated stop segment was misrecognized as such by the first pass. To get a sense of how well the VOT estimation algorithm would perform in the ideal case with a relatively error free first pass, we obtained an HMM segmentation by aligning the speech signal with a known transcription via HMM models for each of the phonemes. The distributions of the VOTs obtained by applying *Algorithm 3* is shown in fig. 5.

Notice the trimodal distribution of the VOTs. The outer modes correspond to gross under and over estimates, i.e., negative or unreasonably high VOT estimates. This occurs about 20 % of the time suggesting that the VOT estimation algorithm provides a reasonably accurate estimate about 80 % of the time when used with an accurate first pass system. In the next section, we present recognition results when the VOT algorithms are used with our baseline HMM as a first pass on test sentences.

5. RECOGNITION EXPERIMENTS

We have developed a two pass framework for recognition. In the first pass, the baseline HMM recognizer, described in section 2, 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 aim to

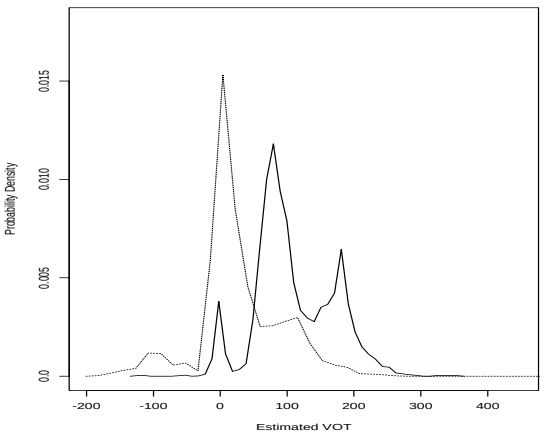

Figure 5: Distributions of VOTs for "t" (dotted) and "d" (solid). Obtained by running the second pass VOT extraction algorithm on a first pass that corresponds to a forced alignment of the HMM models with the true identity.

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 alphabet recognition with a second pass correcting only the confusions shown in table 1 . Due to the asymmetry in the confusion pair statistics, i.e., voiced → unvoiced confusions are considerably higher than the other way around, we targeted only those segments that were classified as an unvoiced stop by the baseline HMM. Thus segments classified as "T","P" or "K" were reanalyzed. The second pass obtained VOT estimates using each of the three algorithms described earlier. Using the VOT estimates, the "T" was reclassified as a "T" only if VOT was greater than 40 ms ("D" if less than 40 but greater than 2 and "E" otherwise); "P" was reclassified as a "B" only if VOT was greater than 30 ms ("B" if less than 30 but greater than 2 ms; "E" otherwise); "K" was reclassified as "K" only if VOT was greater than 50 ms ("A" otherwise).

Notice that by reclassifying in the manner described above, we change only the subset of the confusion matrix displayed in table 1. Furthermore, for our purposes, since we are targeting only those features related to the burst and voicing, it is meaningful to consider only confusions between the three separate classes of sounds — unvoiced stops ({P,T,K}); voiced stops {B,D}; and vowels {A,E}. Table 5. shows the new confusions between these classes.

We see that all the automatic reclassification algorithms reduce the error rate along the voicing dimension by a significant amount (Algorithms 1, 2 and 3 by 18%, 19%, and 53% respectively. Clearly, algorithm 3 is vastly superior to all others in its performance. It should be pointed out, though, that algorithms 1 and 2 were designed for burst detection without a first pass while algorithm 3 was specially designed for our purposes.

	{P,T,K}	{B, D}	{A,E}
{P,T,K}		**67**(b) ; 284(1) **120**(3); 286(2)	**12**(b); 377(1) **106**(3); 407(2)
{B,D}	**646**(b); 284(1) **223**(3); 286(2)		**14**(b); 168(1) **73**(3); 177(2)
{A,E}	**2285**(b); 1078(1) **1008**(3); 1072(2)	**290**(b); 511(1) **603**(3); 475(2)	

Table 2: Number of confusions between broad classes of alphabets separated by burst and voicing. The numbers in brackets indicate type of algorithm used: (b) baseline, and (1),(2),(3) are Algorithms 1,2,3 in second pass mode. Baseline and best algorithms are shown in bold.

6. CONCLUSIONS

We have investigated the possibility of using acoustic-phonetic features for recognition in a two pass manner. In particular, we have focused on the voiced/voiceless distinction between stop consonants in the context of an alphabet recognition task.

We have described our acoustic-phonetic studies that demonstrate that the VOT serves as a stronger dimension of separability than traditional spectral measures like energy banks or cepstra. We have described several different algorithms for automatically extracting the VOT and their use in a continuous speech recognition system. We have compared these algorithms and shown that all of them help improve recognition accuracy and the best among them decreases error rates by as much as 53 percent.

7. REFERENCES

1. Abramson, A.S. and Lisker, L. "Discriminability along the voicing continuum: cross-language tests." *Proceedings of the Sixth International Congress of Phonetic Sciences*, Prague:Academia, 1970,569-573.
2. Fanty, M. and Cole, R. "Speaker-Independent English Alphabet Recognition: Experiments with the E-set," pp.1361-1364, ICSLP,Japan, 1990.
3. L. Djezzar and J. P. Haton. "Exploiting Acoustic-Phonetic Knowledge and Neural Networks for Stop Recognition," pp. 2217-2220, Eurospeech, 1995.
4. 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.
5. C-H. Lee, B-H. Juang, W. Chou, and J. J. Molina-Perez. "A Study on Task Independent Subword Selection and Modeling for Speech Recognition," Proceedings of ICSLP, pp. 1816-1819, 1996.
6. P. Niyogi and P. Ramesh. "Incorporating Voice Onset Time to Improve Letter Recognition Accuracies," Proceedings of ICASSP, 1998.
7. D. Talkin. "A Robust Algorithm for Pitch Tracking (RAPT)," pp. 495-518, in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal (ed)., Elsevier, 1995.
8. Zue, V. W. "Duration of English Stops in prestressed position," Meeting of ASA, Vol.57 (S1),S34(A),1975.