CONSONANT DISCRIMINATION IN ELICITED AND SPONTANEOUS SPEECH: A CASE FOR SIGNAL-ADAPTIVE FRONT ENDS IN ASR*

Kemal Sönmez[1] *Madelaine Plauché*[1,2] *Elizabeth Shriberg*[1] *Horacio Franco*[1]

[1]Speech Technology and Research Laboratory
SRI International, Menlo Park, CA, USA
http://www.speech.sri.com/

[2]Department of Linguistics
University of California, Berkeley, CA, USA
http://www.linguistics.berkeley.edu/

ABSTRACT

The constant frame length in typical ASR front ends is too long to capture transient phenomena in speech, such as stop bursts. However, current HMM systems have consistently outperformed systems based solely on non-uniform units. This work investigates an approach to "add back" such transient information to a speech recognizer, without losing the robustness of the standard acoustic models. We demonstrate a set of phonetically-motivated acoustic features that discriminate a preliminary test set of highly ambiguous voiceless stops in CV contexts. The features are automatically computed from data that had been hand-marked for consonant burst location and voicing onset (extension to automatic marking is also proposed). Two corpora are processed using a parallel set of features: conversational speech over the telephone (Switchboard), and a corpus of carefully elicited speech. The latter provides an upper bound on discrimination, and allows for comparison of feature usage across speaking style. We explore data-driven approaches to obtaining variable-length time-localized features compatible with an HMM statistical framework. We also suggest techniques for extension to automatic annotation of burst location, for computation of features at such points, and for augmentation of an HMM system with the added information.

1. INTRODUCTION

Modeling of speech with hidden Markov models (HMMs) implies a constant rate of information accumulation. Frames of a fixed length are scored uniformly to compute the likelihood that a given utterance is produced by the model. The common fixed frame length of $\sim$25 ms is the time-frequency trade-off in the speech representation. It is well known that such a frame length is too long for capturing information-bearing transient phenomena which may have durations as short as a couple of milliseconds. At the same time, stationary segments, such as vowels, have constant spectral characteristics for much longer regions, on the order of 100 ms. These observations motivate exploring techniques that can provide variable temporal resolution depending on the type of event. This work explores data-driven approaches to such front end adaptation for use within the standard HMM framework.

Approaches based on non-uniform frame lengths have been explored in numerous previous studies. For example, beginning in the 1970s, knowledge-based approaches to speech recognition developed classification systems based on acoustic-phonetic rules [12, 13, 4]. An advantage of such approaches was that the acoustic characteristics for phone discrimination were not limited in resolution. However, performance did not reach that of HMM-based systems using less sophisticated information and a fixed frame length. More recently, segment-based systems [5, 8, 1] address the problem of a constant frame length by representing phone segments using a single feature vector—regardless of segment duration. This approach allows for the use of heterogeneous, phone-class-specific features that focus on phonetically relevant information for discriminating among the confusable sounds within a phone class [7, 10]. In spite of these advantages, however, segment-based systems alone have not been able to outperform state-of-the-art HMM-based systems.

This work aims at combining the advantages of both segmental and HMM systems, by using the HMM system to produce N-best hypotheses with phonetic segmentations. Based on the HMM segmentations, we compute additional phone-specific segment-based features to improve the discrimination of confusable phone classes. Probability models for the additional features are trained from segmentations of training data. For recognition, probability scores for each recognition hypothesis in the N-best list are combined with standard HMM likelihood scores. We demonstrate a set of linguistically motivated features, based on non-uniform front-end extraction units, that successfully discriminate a preliminary test set of voiceless consonants in consonant-vowel (CV) contexts. The features are automatically computed from an extraction region carefully hand-marked by a linguist for burst location. The annotations and feature extraction are applied to two parallel databases: (1) spontaneous-conversational speech from the Switchboard corpus [6] and (2) a corpus of carefully-elicited speech [9]. Inclusion of the latter corpus provides an upper bound on discrimination, and allows us to examine differences in speech style and channel quality while feature definition and extraction is held constant in the parallel corpora. We propose techniques for automatic location of such points in the waveform, computation of features at such points, and augmentation of an HMM system with such information.

The paper is organized as follows: In Section 2, we describe the phone classification task and the database. Features are introduced in Section 3, and the resulting statistics from the proposed features

*A longer version of this work appears in the Proceedings of the NIST Speech Transcription Workshop, College Park, MD, 2000.

on the elicited and spontaneous speech databases are detailed in Section 4. Section 5 describes the decision tree classification of the stops in vocalic contexts via the set of proposed features. Finally, the approaches are discussed from the perspective of automatic speech recognition in Section 6.

2. TASK AND DATABASE

As a first, tractable task in this work, we chose the classification of voiceless unaspirated stops (/p/, /t/, and /k/) in a CV context. We also included /ch/ for comparison purposes. Acoustic information relevant to the identification of stops resides in formant transitions, duration of closure and release of the stop burst, and also in more transient phenomena such as the shape of the spectrum at the burst and the presence or absence of multiple bursts [11]. In certain vocalic contexts, for example, preceding the high front vowel, /i/, long term cues may be neutralized, resulting in a dependence on the transient phenomena for the identification of stops, and a corresponding increase in confusion rates for both humans and machines [9]. The set of voiceless stops in CV tokens presents a challenge to automatic processing approaches that average transient information of stops over many frames, and thus proves to be a good starting point for localized feature modeling.

3. FEATURES OF A CV TOKEN

For the purposes of this cross-corpus study, we considered the following subset of acoustic features known to be important cues in the identification of stop place [11]: (1) voice onset time (VOT), (2) multiplicity of bursts, and (3) gross shape of burst spectrum. VOT is the duration of time the vocal cords take to begin periodic vibration after the release of a consonant. The predicted order of VOT averages, derived from their articulation and manner, for voiceless stops and the affricate /ch/ are: /p/, /t/, /k/, /ch/.

Figure 1 shows the distribution of VOT and multiplicity of bursts for elicited and spontaneous speech. The bar graphs represent values averaged over all speakers and all vocalic contexts. VOT, as predicted, is a strong function of stop identity for both elicited and spontaneous speech. Multiplicity of bursts, however, serves as a useful discriminant only for the elicited database, as the articulation of stops in spontaneous speech may have a faster release overall, and thus velar stops may be less prone to the phenomena of multiple bursts.

The constriction of the articulation of a voiceless stop and its release generate distinctive spectral characteristics at the burst that are somewhat invariant across different vocalic contexts. Stop burst spectra for labials (/p/), alveolars, (/t/), and velars (/k/) have been described as “diffuse-falling” (majority of energy in the low frequency region), “diffuse-rising” (majority of energy in the high frequency region), and “compact” (peak of energy in the mid frequency region) [2]. Figures 2, 3 and 4 show examples of the spectrum at the burst, as well as linear and piecewise linear fits to the spectrum, for /pa/, /ta/, and /ka/, respectively. Derived features include the slopes of the linear and piecewise linear fits, the mean squared error of the fits, and the location in the frequency range of the node for the piecewise linear fit. The last feature is particularly helpful for distinguishing /t/ from /k/ bursts.

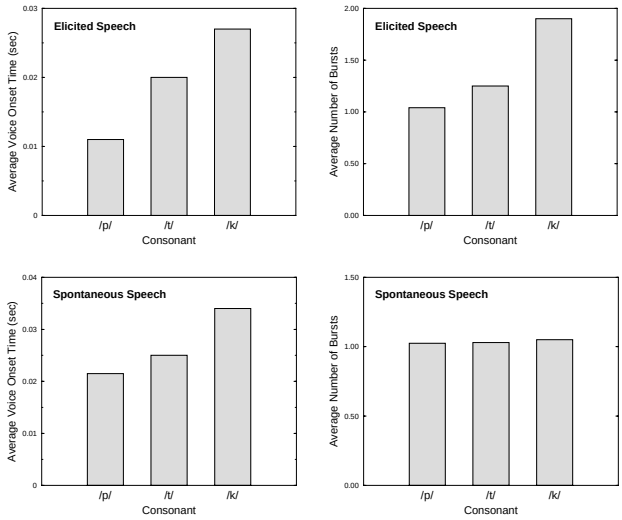

Figure 1: Comparison of average VOT and average number of bursts for elicited and spontaneous speech. Values are averaged over speakers and over vocalic context.

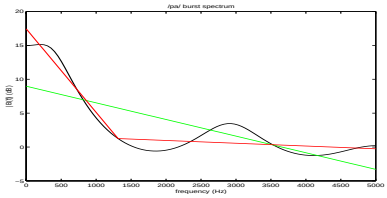

Figure 2: Spectrum, linear, and piecewise linear fits for /pa/. The “diffuse-falling” shape of the spectrum is captured by the negative slope of the linear fits. Note that the node lies below 2000 Hz.

4. DECISION TREES

In this section we describe the analysis and visualization of the proposed set of features via decision trees. For the classification problem over the set (/p/, /t/, /k/, /ch/), we train CART-style decision trees, as in Figure 5. In Table 1, we show the classification performance of the decision tree on a test set for elicited speech. The corresponding performance summary on spontaneous speech is given in Table 2. The classification accuracies are around 84%.

We also rank and compare the usage of the features shown here as well as formant transition information across elicited and sponta-

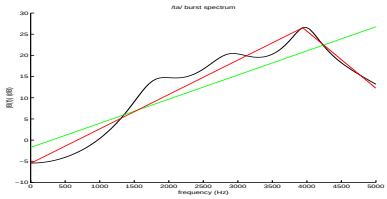

Figure 3: Spectrum, linear, and piecewise linear fits for /ta/. The “diffuse-rising” shape of the spectrum is captured by the positive slope of the linear fits. Note that the node lies above 4000 Hz.

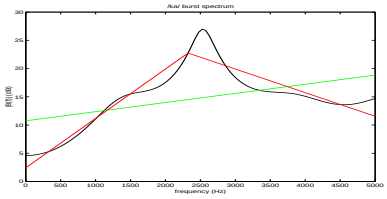

Figure 4: Spectrum, linear, and piecewise linear fits for /ka/. The prominent peak of the spectrum is captured by a high mean squared error of the linear fit and by the mid-frequency location (2000–4000 Hz) of the fitted node.

Elicited speech database, accuracy = 84.11% (905/1076)

	P	T	K	CH	TOTAL	CORR
P	243	16	9	1	269	243
T	37	180	49	3	269	180
K	12	37	213	7	269	213
CH	0	0	0	269	269	269

Table 1: Decision tree classification of stops in vocalic contexts for elicited speech.

neous databases. Table 3 shows the frequency of usage of features in tree classification. VOT is the most prominent feature on both elicited and spontaneous databases. The tree in Figure 5 shows that VOT is especially helpful in classifying /ch/ and /p/. Burst multiplicity, as previously mentioned, is only useful in the elicited database, where it is used to classify velars from other stops. The tree also contains information on formant transitions into the following vowel, which is found to be useful in both databases; here it picks out labials (which have characteristically low formant onsets at the release of the burst) from other stops. The node frequency of the piecewise linear fit is also a consistently used feature in both the elicited and spontaneous databases; here it functions to distinguish velars from alveolars.

5. ASR PERSPECTIVE

We have shown that hand-labeled acoustic events, some of which are temporally localized, provide features with rich information content for the classification of easily confused phones. Here, we discuss the issues in extending such an approach to ASR systems. The focus is on automatic location of information-bearing points in the waveform and statistical extraction of localized features. Another fundamental question is the determination of the best way to augment or modify current HMM systems to use such information.

Spontaneous speech, accuracy = 83.57% (234/280)

	P	T	K	CH	TOTAL	CORR
P	61	6	3	0	70	61
T	12	51	6	1	70	51
K	8	2	52	8	70	52
CH	0	0	0	70	70	70

Table 2: Decision tree classification of stops in vocalic contexts for spontaneous speech.

Elicited speech		Spontaneous speech	
feature	**usage**	**feature**	**usage**
VOT	0.57	VOT	0.54
f_2	0.19	f_2 slope	0.29
number of bursts	0.13	f_2	0.12
burst node freq	0.10	burst node freq	0.05
f_2 slope	0.01	number of bursts	0.00

Table 3: Decision tree usage of features.

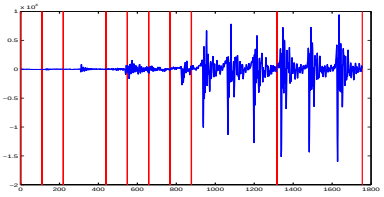

Figure 6: Adaptive frame length analysis by the best basis algorithm.

As an important example of the signal processing issues involved, we demonstrate automatic localization of the voicing onset.

We have made use of the best basis algorithm [3] in segmenting transient and stationary speech segments by an adaptive frame length front end. In this framework, automatic voicing onset location is carried out by temporal segmentation into varying-length frames depending on the stationarity of the underlying signal segment (Figure 6). This type of front-end processing may also be suitable for burst localization.

Finally, we discuss possible ways of augmenting an HMM system with localized features. One straightforward way of augmentation is via N-best list rescoring from alignments as shown in Figure 7. The CV context is bracketed by alignments; subsequently, the features obtained from the CV are scored and used as an additional knowledge source in rescoring of the N-best list.

6. SUMMARY AND FUTURE WORK

This work has explored data-driven approaches to temporal front end adaptation. We have carried out statistical extraction and characterization of useful time-localized features obtained from data hand-labeled for relevant events. Such work constitutes a first step toward demonstrating the discrimination power of localized features on a classification task, for both careful and spontaneous speech. We have also discussed signal processing techniques to automate the accurate localization of information-bearing events, and possible methods of augmentation or modification of current HMM systems to use localized features as side information.

7. ACKNOWLEDGMENTS

This work was funded by the Department of Defense, and by NSF-STIMULATE Grant IRI-9619921. The views herein are those of the authors and do not necessarily reflect those of the funding agencies.

8. REFERENCES

1. Bacchiani, M., and Ostendorf, M. (1998). Using automatically-

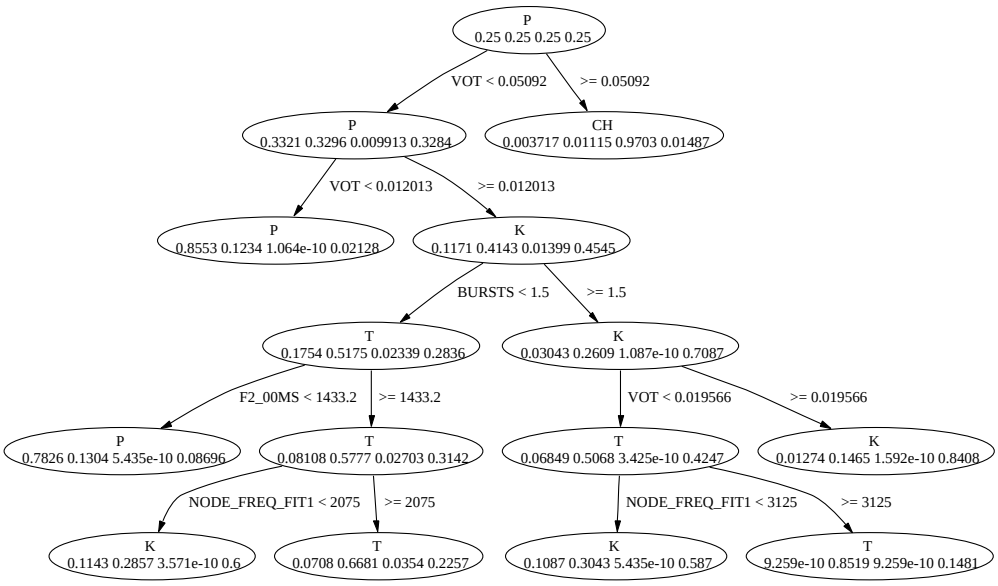

Figure 5: Decision tree for classification of stops in vocalic contexts.

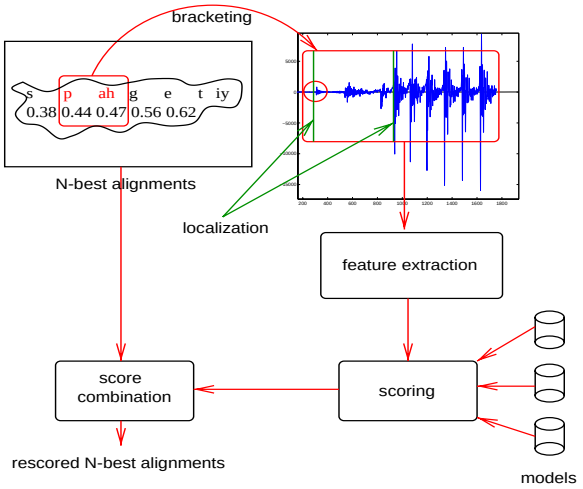

Figure 7: HMM system augmentation with localized feature modeling.

derived acoustic subword units in large vocabulary speech recognition. *Proc. International Conference on Spoken Language Processing* Vol. 5, pp. 1843–1846.

2. Blumstein, S., and Stevens, K. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America*, 66(4), 1001–1017.

3. Coifman, R. R., and Wickerhauser, N. V. (1992). Entropy-based algorithms for best basis selection. *IEEE Trans. Inform. Th.* 38(2), 713–718.

4. Cole, R., Stern, R., and Lasry, M. (1986). Performing fine phonetic distinctions: templates versus features. In J. S. Perkell and D. M. Klatt (eds). *Variability and Invariance in Speech Processes*, Erlbaum: Hillsdate, N.J.

5. Glass, J., Chang, J., and McCandless, M. (1996). A probabilistic framework for feature-based speech recognition. *Proc. International Conference on Spoken Language Processing*, pp. 2277–2280, Philadelphia, PA.

6. Godfrey, J., Holliman, E., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *Proc. ICASSP*, Vol. 1, pp. 517–520.

7. Halberstadt, A., and Glass, J. (1997). Heterogeneous measurements for phonetic classification. *Proc. Eurospeech*, pp. 401–404, Rhodes, Greece.

8. Ostendorf, M., Digalakis, V., and Kimball, O. (1996). From HMMs to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Trans. Speech and Audio Processing*, 4(5), 360–378.

9. Plauché, M., and Sönmez, K. (2000). Machine learning techniques for the identification of cues for stop place. To appear in *Proc. International Conference on Spoken Language Processing*, Beijing.

10. Schmid, P. (1996). Explicit N-best Formant Features for Segment-Based Speech Recognition. Ph.D. thesis, Department of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, Portland.

11. Stevens, K. (1999). *Acoustic Phonetics*. Kluwer Academic Publishers, Boston.

12. Weinstein, C., McCandless, S., Mondshein, L. and Zue, V. (1975). A system for acoustic-phonetic analysis of continuous speech. *IEEE Trans. Acoust. Speech Signal Process.*, 23, 54–67.

13. Zue, V. (1985). The use of speech knowledge in automatic speech recognition. *Proc. IEEE*, 73, 1602–1615.