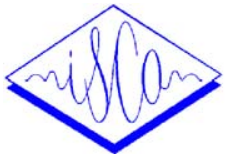


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

AUTOMATIC DETECTION OF MANNER EVENTS BASED ON TEMPORAL PARAMETERS

Ariel Salomon and Carol Espy-Wilson
Electrical and Computer Engineering Department, Boston University
8 St. Mary's St., Boston, MA, USA 02215
ariel@bu.edu, espy@bu.edu

Abstract

In this study, we investigated how well acoustic events extracted from a cross-spectral temporal measure could be used to classify the manner and voicing of consonants. In particular, we developed seven measures that look at the strength and time difference between various onsets and offsets of acoustic energy. Consistent with findings by Shannon et al. (1995), our classification results show that manner and voicing information can be determined from dynamic temporal cues.

Keywords: landmark detection, acoustic events, knowledge-based speech recognition, phonological features

1. INTRODUCTION

This research continues work toward the development of a knowledge-based representation of a speech signal. The approach taken is based on the theory that speech segments are collections of articulatorily-motivated phonological features [1], with canonical acoustic properties. Modeling a speech signal as a sequence of alternations of phonological features leads to an approach to speech recognition involving location of landmarks in the signal corresponding to changes in manner features, and searching for other types of information in the vicinity of these landmarks.

Previous work [2,3] has undertaken the task of detecting manner events in a speech signal using a variety of spectral and temporal parameters. One particular parameter that has often been used is a first difference of spectral energy, especially for stop detection. This paper discusses the development of a classification system for consonants that is based on temporal measures derived from a cross-spectral first difference parameter.

This work is motivated in part by the findings in [4] which showed that humans are able to identify the manner and voicing of consonants occurring in spectrally-impoverished speech using primarily temporal cues.

2. METHOD

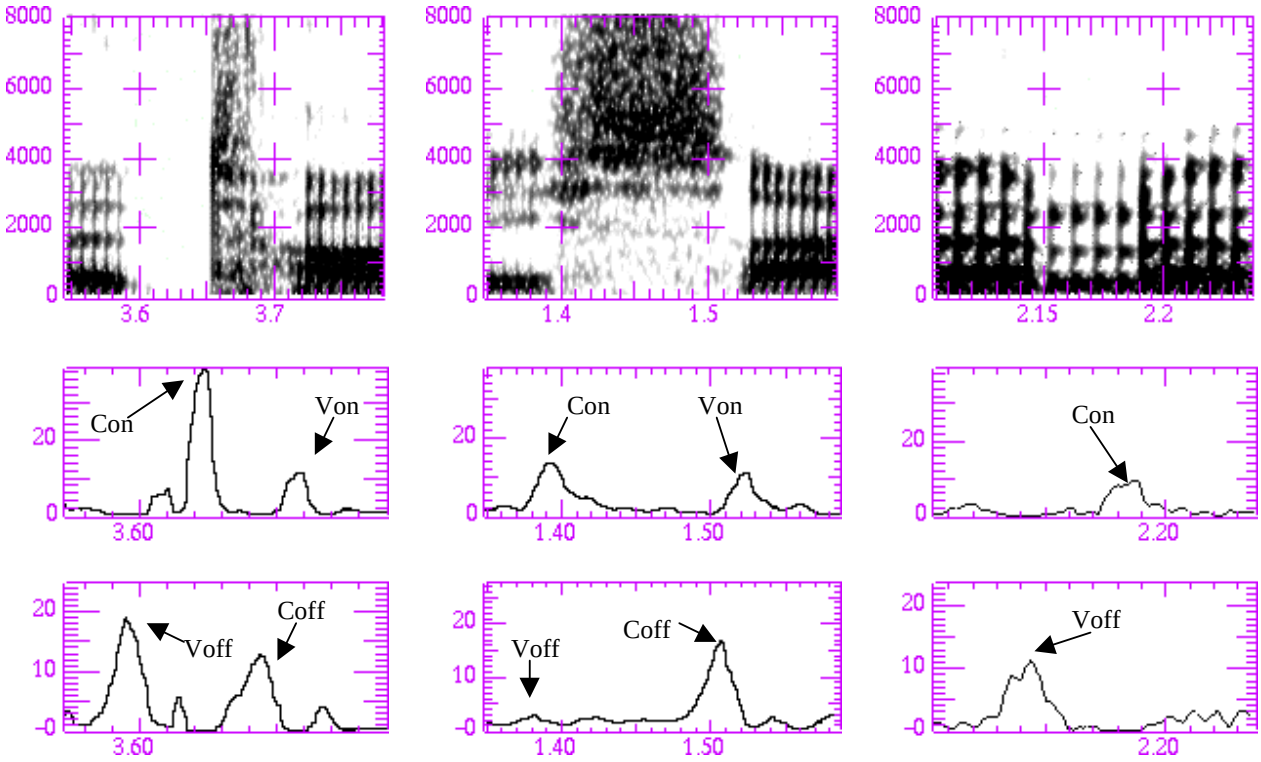
2.1 Database

In this study, we focus on singleton consonants occurring in an intervocalic context. To develop the temporal measures to distinguish between the consonants, we used 626 SI and SX sentences from the TIMIT training set. For classification, we used 382 SI and SX sentences from the TIMIT test set.

2.2 Onsets and offsets

All of the temporal parameters developed are based on a cross-spectral average of onset and offset in energy. This is similar to the 'burst' parameter used by [2] except that it involved computing the first difference between two averaged FFT windows, rather than individual FFTs, and also was computed with a longer first difference time period.

To compute the parameters, first a 6ms Hamming windowed FFT was computed in 1ms intervals. Next, spectra within a 15ms window were averaged. The first difference was computed individually in each of 257 spectral channels between averaged spectra spaced 18ms apart. The onsets (positive first differences) and offsets (negative first differences) were added up separately to produce two values at each time step. The onset and offset values were divided by the number of spectral bins to produce measures of average onset and offset in dB.



(a) intervocalic stop **(b) intervocalic fricative** **(c) intervocalic nasal**

Fig 1. Spectrogram and onset and offset parameters for a stop, fricative and nasal occurring in intervocalic contexts.

An example of typical onset and offset patterns for a stop, fricative and nasal occurring in intervocalic positions are shown in Fig. 1. (Note that the offset parameter has been inverted.) In the case of the stop and fricative, four events occur: (a) the offset of the preceding vowel (Voff), (b) the onset of sound during the consonant (Con), (c) the offset of the consonant (Coff) and (d) the onset of the following vowel (Von). However, the pattern of events is different. The time difference between Con and Voff is much larger for the stop due to the complete closure formed. Relatedly, the time difference between Coff and Con is longer for a fricative since noise is generated throughout its production. Unlike the obstruents, in the nasal (and other sonorant consonants), there are only two events, the offset of energy during the transition from the vowel to the nasal and the onset of energy during the transition from the nasal to the following vowel. To fit the template for the obstruent pattern, we call the onset Con and the offset Voff.

2.3 Parameters

To find the relevant events, a simple peak detector was run on the onset and offset parameters. The peaks were searched for in a region defined by the midpoints of the adjacent vowels. First, the largest onset (Con) was found. Then, the largest offset prior to Con (Voff) and the largest offset and onset following Con (Coff and Von, respectively) were located. Note that in the case of the sonorant consonants, peaks corresponding to Coff and Von won't typically be found. Peaks were extracted with a minimum separation in time (15ms) and a minimum strength (0.1dB). Peaks were also combined if the maximum dip between them was small, above 0.7x the height of the larger peak.

To capture the difference in the patterns we observe for the consonants, three temporal measures were computed. The first parameter computes the time difference between Con and Voff. In the case of stops, this duration corresponds to the stop closure. In the case of fricatives, this duration will be small (typically close to zero) since the beginning of frication is usually coincident with the end of the vowel. In the case of the voiced fricatives, where the vocal folds continue to vibrate, this time difference may actually be negative since overlap in the articulatory gestures is more likely. In the case of sonorant consonants where there is only one onset

and one offset, this difference will measure the duration of the sonorant consonant.

The second parameter measures the time difference between Coff and Con. This difference should correspond to the duration of the burst for stops, and the duration of the frication noise for fricatives; this measure is not relevant for the sonorant consonants.

The last temporal parameter measures the time difference between Von and Con. In the case of stops, this duration should correspond to the voice onset time. In the case of fricatives, this measure is similar to the one above, corresponding primarily to the duration of the frication noise. As before, this measure is irrelevant for sonorant consonants.

Finally, in addition to computing the difference in time between the various onsets and offsets, the strengths of the Voff, Con, Coff and Von peaks are computed. Taken together, these measures result in a 7 dimensional vector.

Note that because this set of peaks was extracted for all consonants, it is possible and in fact likely that some of the peaks may not be found in all cases.

2.4. Classifier

The vector of parameters described in the previous section was used to make several types of decisions including sonorant vs. nonsonorant; among sonorants, nasal vs. non-nasal (i.e. semivowel); among obstruents, continuant vs. noncontinuant where affricates were considered noncontinuant; voicing for fricatives; stops vs. affricates; and voicing for stops.

The decision test used was a Gaussian Likelihood Ratio Test with means and covariances computed over the subset of the data for each class that had valid values (i.e. the peaks were found):

$$(y-m_0)\Sigma_0^{-1}(y-m_0) - (y-m_1)\Sigma_1^{-1}(y-m_1) > \Gamma$$

The Γ threshold parameter was adjusted to minimize Pr[Type I Error] + Pr[Type II Error], which worked better than strictly minimizing Pr[Error] with respect to confusions, and should be more robust to changes in relative consonant frequencies.

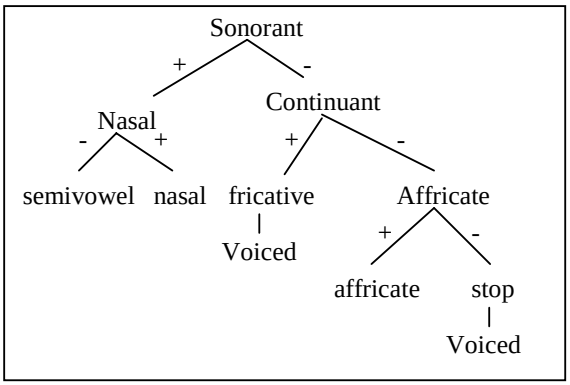


Fig. 2. Decision rule hierarchy.

In cases where a measure was missing (due to a peak not found) for a particular consonant, it was classified based only on the existing measures. It is likely that some increased error was caused by not considering conditional probabilities in these cases, since nonexistence of a measure can sometimes be a strong cue.

All of the analyzed consonants were classified by hierarchically using the rules as shown in Fig. 2.

3. RESULTS

The results, in the form of a confusion matrix, are shown in Tables 1 and 2 for the training and test data, respectively. Correct manner classification was 82% (training) and 88% (test) in the case of stops, 74% (training) and 84% (test) in the case of affricates, 47% (training) and 43% (test) for fricatives, 71% (training) and 77% (test) for nasals, and 76% (training) and 66% (test) for the semivowels.

The poorest manner results were for fricatives. Note that most of the misclassified fricatives were called sonorant consonants. In the case of voiced fricatives, this sort of classification may be due to lenition where the fricative is actually realized as a sonorant consonant [5]. For both voiced and unvoiced fricatives, this misclassification should improve with sonorant measures that depend on frequency such as those computed in [2] and [3].

Of the stops classified as stops, the voiced ones are correctly classified as voiced with an accuracy of 88% (training and test). The unvoiced stops were classified as unvoiced with an accuracy of 68% (training) and 53% (test). Of the fricatives

Table 1. Confusion matrix for training data VS=voiced stop, US=unvoiced stop, fl=flap (underlying stop), Af=affricate, VF=voiced fricative, UF=unvoiced fricative, Ns=nasal, Sm=semivowel (glides and liquids), Tot=total #

	Classified Manner Class (training)							
Label	**VS**	**US**	**Af**	**VF**	**UF**	**Ns**	**Sm**	**Tot**
VS	130	18	11	0	0	12	2	173
US	60	127	32	0	4	11	0	234
fl	28	4	11	2	2	86	33	166
Af	8	3	63	3	5	2	1	85
VF	6	0	1	41	25	80	66	219
UF	8	1	1	10	141	49	37	247
Ns	5	1	0	3	1	224	80	314
Sm	0	0	1	1	4	64	222	292

Table 2. Confusion matrix for test data VS=voiced stop, US=unvoiced stop, fl=flap (underlying stop), Af=affricate, VF=voiced fricative, UF=unvoiced fricative, Ns=nasal, Sm=semivowel (glides and liquids), Tot=total #

	Classified Manner Class (test)							
Label	**VS**	**US**	**Af**	**VF**	**UF**	**Ns**	**Sm**	**Tot**
VS	99	13	3	0	0	6	1	122
US	65	73	17	0	2	2	2	161
fl	14	2	3	2	0	57	15	93
Af	0	1	21	1	2	0	0	25
VF	3	2	1	14	7	44	54	125
UF	1	0	1	10	82	25	19	138
Ns	4	0	1	5	1	147	33	191
Sm	1	1	0	1	2	68	142	215

classified as continuants, the voiced ones were correctly classified as voiced with an accuracy of 62% (training) and 67% (test). Similarly, the unvoiced fricatives were classified as unvoiced with an accuracy of 93% (training) and 89% (test).

Additional experiments were performed to determine how well each individual measure could make each of the distinctions, and the two best measures for each decision. There were some interesting patterns with respect to the measures that were most useful in making distinctions. One major example of this is the fact that the strength of the Con peak was one of the two most important measures for all of the top-level manner decisions: sonorant vs. non-sonorant, semivowel vs. nasal, and continuant vs. noncontinuant. This is most likely due to both the robustness of this parameter, as it should exist for all input segments, as well as the fact that overall strength of spectral discontinuities is a major cue used in understanding speech.

Another interesting pattern is that the two most important features distinguishing sonorants from obstruents were Con strength and Coff strength. This is logical because obstruents are characterized by a noisy section between Con and Coff, whereas sonorants are not expected to have a Coff event. A large proportion of the fricatives had relatively weak or missing Coff events, which probably contributed to their misclassification.

4. CONCLUSIONS

Clearly the classification results show that manner and voicing information is encoded in the temporal structure of consonants. At present we are using a single temporal measure from which we extract several features. In future work, we will explore additional temporal cues and evaluate their performance on spectrally-impoverished speech. The other major direction for this work is to build a full recognition system that includes use of temporal information for detection of consonants. This would include recognition of consonant clusters, and vowel landmark detection.

5. ACKNOWLEDGEMENTS

This research was supported in part by NIH grant 1-K02-DC00149-01A1 and NSF grant SBR-9729688.

6. REFERENCES

1. K. N. Stevens and S. J. Keyser, “Feature geometry and the vocal tract”. *Phonology* 11 (1994) 207-236.
2. N. N. Bitar and C. Y. Espy-Wilson, “Knowledge-Based Parameters for HMM Speech Recognition”. *Proceedings, IEEE Conf. ASSP* (1996) 29-32.
3. S. Liu, “Landmark detection for distinctive feature-based speech recognition”. *J. Acoust. Soc. Am.* 100 (1996) 3417-3430.
4. R. V. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, “Speech Recognition with Primarily Temporal Cues”. *Science* 270 (1995) 303-304.
5. C. Y. Espy-Wilson, “Acoustic measures for linguistic features distinguishing the semivowels /w j r l/ in American English”. *J. Acoust. Soc. Am.* 92 (1992) 736-757.