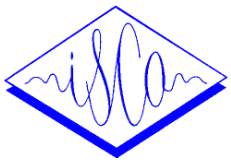

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

AUTOMATIC CONTEXT-SENSITIVE MEASUREMENT OF THE ACOUSTIC CORRELATES OF DISTINCTIVE FEATURES AT LANDMARKS

Mark Johnson

Research Laboratory of Electronics, MIT
Cambridge, MA 02139, USA

Abstract

This paper models speech recognition as the estimation of distinctive feature values at articulatory landmarks [8]. Toward this end, we propose modeling each distinctive feature as a table containing phonetic contexts, a list of signal measurements (acoustic correlates) which provide information about the feature in each context, and, for each context, a statistical model for evaluating the feature given the measurements.

The model of a distinctive feature may include several sets of acoustic correlates, each indexed by a different set of context features. Context features are typically lower-level features of the same segment, e.g. manner features ([continuant, sonorant]) provide context for the identification of articulator-bound features ([lips, blade]). The acoustic correlates of a feature can be any static or dynamic spectral measurements defined relative to the time of the landmark. The statistical model is a simple N-dimensional Gaussian hypothesis test.

A measurement program has been developed to test the usefulness of user-defined acoustic correlates in user-defined phonetic contexts. Measures of voice onset time and formant locus classification are presented as examples.

1 Recognition of planning units

When a word changes tense, case, or number, the phonemes in the word often change in predictable ways. Distinctive feature notation was developed as a compact notation for these empirical rules of phonological alternation. As such, distinctive features are a model of a data representation used by the human brain to plan speech production [2].

In order to be useful for modeling speech production, a set of distinctive features must be as compact as possible, while providing unique representations for all distinguishable allophones. In English, for example, there are about forty phonemes, and one to two hundred allophones, all of which may be uniquely represented in terms of about twenty binary distinctive features. In this paper, we will use a binary articulatory feature set capable of representing some syllabic and prosodic information, in addition to traditional allophone distinctions (figure 1).

Since there are fewer distinctive features than phonemes, and since they provide a more parsimonious representation of coarticulation, automatic recognition of distinctive features should be more efficient than automatic recognition of phonemes. Meng et al. were able to reduce complexity, at a slight cost in performance, by using distinctive features as an intermediate data representation in a neural network vowel recognizer [5]. The system of Deng et al. models coarticulation as the overlap of distinctive features; using lightly trained statistical models of the resulting allophonic feature bundles, they achieved 72% correct recognition on a subset of TIMIT [1].

time (ms)	1178	1253	1280	1353	1462
symbol	k	uh	k	k	iy
rel. or closure	r		c	r	
Syllabic		+			+
Reduced		-			-
Consonantal	+		+	+	
Continuant	-		-	-	
Sonorant	-		-	-	
Lips					
Blade					
Dorsum	+		+	+	
Round	+	+		-	-
Anterior					
Distributed					
High	+	+	+	+	+
Low	-	-		-	-
Back	+	+		-	-
Adv. Tong. Rt.		-			+
Const. Phar.		-			-
Nasal	-		-	-	
Spread Glottis	+			+	
Const. Glottis	-			-	
Stiff Vocal Folds	+		+	+	

Figure 1: The word "cookie," represented in features taken from an articulatory feature set (sentence "Below the city is a cookie," speaker JW)

The systems of Meng et al. and Deng et al. apply distinctive feature notation to the standard speech recognition model, in which speech is composed of spectrally homogenous temporally sequential micro-segments. In contrast to this, we have been working with a model in which the linguistic content of speech is located at a series of articulatory landmarks, and the speech signal in between landmarks is merely a carrier for coded information about distinctive feature values at the landmarks [8].

Information about distinctive features at a landmark is

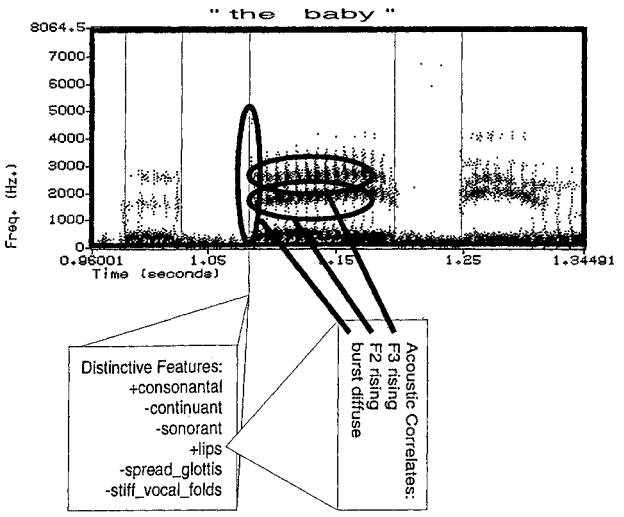

Figure 2: The acoustic correlates of the feature [+lips] at a stop release include formant motions in the following vowel.

coded into the surrounding acoustic signal according to the physical laws of speech production. Place features at a consonant release, for example, have an effect on formant motion in the following vowel, and any statistical model of a place feature should include measurements of the resulting formant motion (figure 2). Spectral measurements at any point in the signal which correlate with the value of a distinctive feature are often called the "acoustic correlates" of the feature.

This paper will define a class of context-dependent acoustic correlates of distinctive features for automatic speech recognition. The acoustic correlates will be a set of signal measurements, known to have a statistical correlation with the value of the distinctive feature, in a phonetic context described by information recorded in the distinctive feature model. Section 2 will describe the coding of phonetic context in terms of context features, and some limitations of this approach. Section 3 will describe the signal measurements which are used as acoustic correlates, including a description of the articulatory landmarks on which they are anchored, and will give some examples. Finally, section 4 will describe training and application of the Gaussian classifier.

2 Context features

Recognizing distinctive features instead of segments both decreases and increases the context dependency of each model. Many of the traditional context dependencies, such as the dependency of formant loci on the neighboring vowel quality, can be almost eliminated by choosing sensible acoustic correlates, as discussed in the next section. The atomic nature of features, however, introduces a whole new set of dependencies: for example, spectral shape of the burst is an important cue for the place of stop releases, but not for the releases of nasals.

For simplicity, the phonetic context of a distinctive feature can be expressed in terms of other distinctive features. These context features may include previously recognized

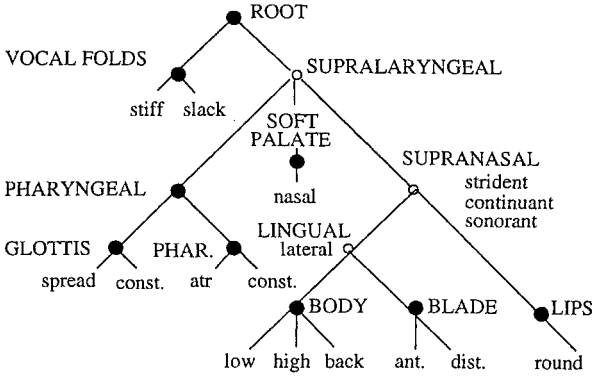

Figure 3: A representation of the Keyser-Stevens articulatory feature geometry

features at the same landmark: for example, the value of the feature [sonorant] can be used as context for classification of the place feature [lips]. The order of feature evaluation in our development system is based on the Keyser-Stevens articulatory feature geometry (figure 3): manner features are evaluated first, followed by features of the subvelar articulators, followed finally by features of the tongue and lips, and identification of the primary articulator [3].

3 Acoustic correlates

3.1 Landmarks

The motions of different articulators are often loosely coordinated, but there seem to be critical times around which these motions must be aligned in order to convey information. For example, in production of a syllable-final nasal, the velopharyngeal port may open at almost any time during the vowel, but if it has not opened by the time the mouth closes, the consonant will not be heard as a nasal.

If there is a primary articulator (e.g. lips, tongue blade, tongue body, glottis), these critical landmark times seem to occur at the moment when the primary articulator reaches its planned target, e.g. the moments of closure and release of a stop, nasal, or fricative, and the moment of maximum constriction of a glide. We can think of each of these landmarks as a flag, marking the production of an allophone, whose identity is determined by the state of the articulators at the time of the landmark. The goal of automatic distinctive feature recognition is to identify the state of the articulators at each landmark.

Landmark insertion and deletion errors can be corrected by a minimum-error lexical access routine, just as segment errors are corrected in a stochastic segment recognizer [6]. The advantage of using landmark candidates instead of segment candidates is that landmarks – consonant closures and glide minima – can be identified more accurately with less computation than segment boundaries. Working only with changes in peak FFT amplitudes, Liu has been able to identify 95% of the obstruent landmarks, with a 4% insertion rate, and 88% of the sonorant consonantal landmarks, with a 26% insertion rate [4]. These deletion rates are comparable to the 5-10success rate like this should be sufficient for a standard Viterbi lexical access algorithm. The insertion rate for sonorant landmarks is quite high, but most of these

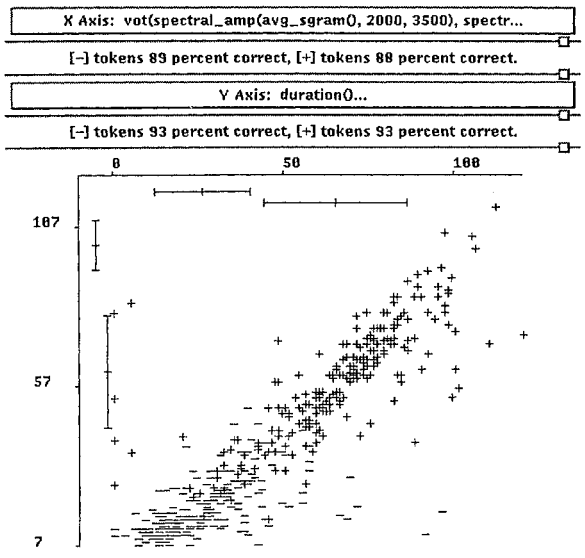

Figure 4: Scatter plot of [+spread_glottis] and [-spread_glottis] stop release tokens: distance between discovered burst and discovered voice onset (X axis), plotted against stop release duration transcribed in TIMIT (Y axis).

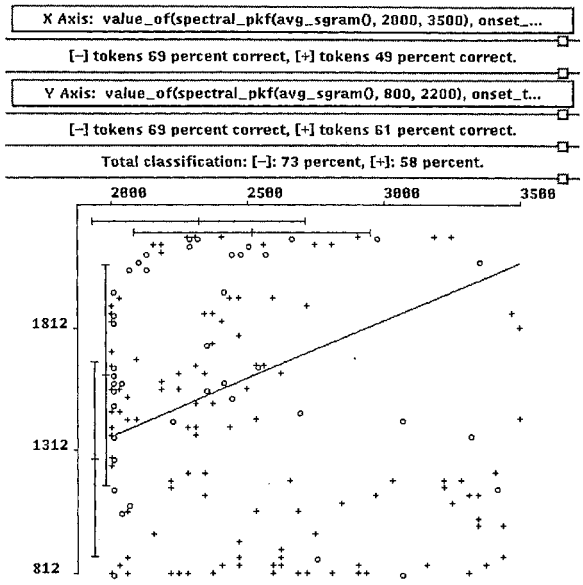

Figure 5: Scatter plot of labial ('+') and velar ('o') voiced stop releases: measures of second formant (Y axis) and third formant (X axis) at transcribed voice onset.

insertions are caused by glide boundaries, which should be easy to identify with more careful spectral processing.

3.2 Signal measurements

An acoustic correlate of a distinctive feature is any spectral measurement defined relative to the landmark which is correlated with the value of the feature at the landmark. There is an enormous literature in phonetics describing the acoustic correlates of distinctive features; the difficult part about designing a feature recognizer is finding specific algorithms and parameter values which can be used by a computer to measure these acoustic correlates robustly [7].

For example, voice onset time (VOT) is positively correlated with the feature [spread_glottis] at stop release landmarks: stops which are [+spread_glottis] (aspirated) have long voice onset times, and stops which are [-spread_glottis] (unaspirated) generally have short onset times. If we know that a stop release has occurred, we should be able to find the moment of voice onset by looking for the onset (within 9dB of the vowel level) of low frequency energy (0 to 400 Hertz). VOT can be measured as the difference between this time and the time of the burst. The burst can be identified in a quiet environment by looking for a sharp rise in middle or high frequency energy; in noise, we might need a more robust approach.

Figure 4 is a scatter plot of VOT, measured as described above, versus duration of the stop release segment, as transcribed in TIMIT. Plus signs ('+') represent [+spread_glottis] stop releases (unvoiced stops not followed by 'q' or preceded by 's': 227 tokens), and minus signs represent [-spread_glottis] tokens (voiced stops: 230 tokens). These were classified with only 93% accuracy by the transcribed segment duration; casual error analysis indicates that some of the tokens marked [+spread_glottis] were probably not aspirated (e.g. /k/ in "stakeout"), and some of the tokens marked [-spread_glottis] were heavily fricated (e.g. /g/ in "big goat"). The VOT measure described above classified the tokens almost as well as the transcribed duration, and was close to the transcribed duration (within 15 milliseconds) 88% of the time. Several of the 12% measurement errors were failures to find a weak burst, which are plotted at VOT=0 in figure 4.

Figures 5 and 6 demonstrate the usefulness of acoustic correlates defined at times beyond the traditional boundaries of a segment. Figure 5 shows 122 labial ('+') and 49 velar ('o') voiced stop release tokens, plotted according to peak frequency measures in the F2 and F3 bands at voice onset. The peak-picking algorithm fails frequently (recording peaks at the band edges), but even the tokens with reasonable formant values show significant overlap. Figure 6 shows the same tokens, but now the F2 onset frequency (Y axis) is plotted against a measure of F2 at the vowel center, up to 100 milliseconds away (X axis). The classification algorithm is able to use the vowel center F2 as context information for the onset, and by weighting the two and combining them (the line shown has a slope of 0.65), we 74% correct place classification [9].

4 Feature classifier

A recognition model of a distinctive feature contains three lists: a list of contexts, a list of acoustic correlates, and a list of feature classifiers. In order to estimate the value of a feature, the recognizer reads through the list of possible contexts until it finds a match with the current landmark. The corresponding acoustic correlates are measured, and their values are used to classify the feature.

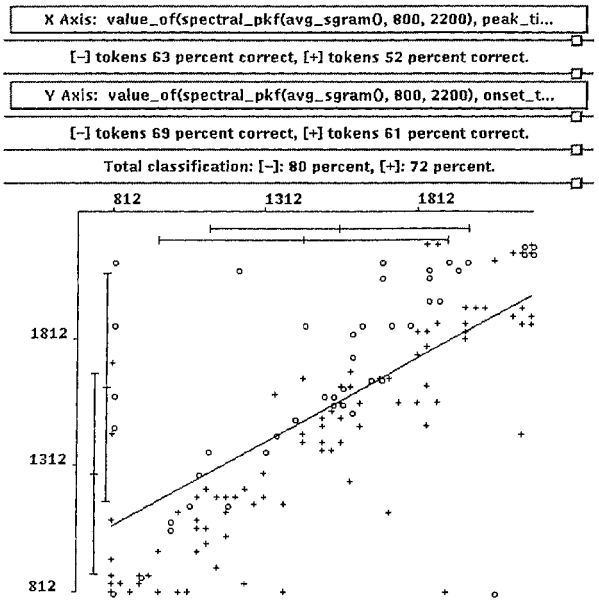

Figure 6: Scatter plot of labial ('+') and velar ('o') voiced stop releases: measures of second formant at vowel onset (Y axis) and vowel peak (X axis).

The current development system uses an N-dimensional Gaussian classifier, which is implemented as a set of weights and an offset. The acoustic measurement values are weighted, summed, and shifted so that zero is the classification threshold. This score can be recorded as a confidence score, or quantized: if the score is positive, the feature is marked "+", otherwise, "-."

The classifier is trained using a measurement program which accepts lists of context features and acoustic correlate definitions, and outputs measurement statistics, including the weights for a Gaussian classifier [7]. The program searches a speech database, transcribed with either phonemes or features, to find landmarks which match the phonetic context, measures the given acoustic correlates at each landmark, and compiles a measurement histogram and statistics. The histogram, classification weights, and individual and collective classification scores can be viewed using the interactive display demonstrated in figures 4 through 6.

5 Conclusion

We have described a framework for developing and testing algorithms to automatically measure the acoustic correlates of distinctive features at landmarks. Only a few of our distinctive feature models have been fleshed out with usable acoustic correlates, but we hope to develop more of these acoustic correlate measurements in the months ahead. When coupled with a landmark detector [4] and a lexical access routine capable of matching recognized features to the existing feature lexicon [8], these feature models will constitute a minimally operational feature-based speech recognizer.

(Research supported in part by the National Science Foundation.)

References

[1] Li Deng and Don X. Sun. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. *JASA*, 95(5):2702–2719, May 1994.

[2] Michael Kenstowicz. *Phonology in Generative Grammar.* Blackwell, Cambridge, Massachusetts, 1994.

[3] S. J. Keyser and K. N. Stevens. Feature geometry and the vocal tract. (submitted to Phonology), 1993.

[4] Sharlene A. Liu. Locating landmarks in utterances for speech recognition. *JASA*, 93:2320, April 1993.

[5] H. M. Meng, V. W. Zue, and H. C. Leung. Signal representation, attribute extraction, and the use of distinctive features for phonetic classification. In *DARPA Speech and Natural Language Workshop*, Pacific Grove, CA, February 1991.

[6] Mari Ostendorf and Salim Roukos. A stochastic segment model for phoneme-based continuous speech recognition. *Trans. ASSP*, 37(12):1857–1869, December 1989.

[7] Michael Philips and Victor Zue. Automatic discovery of acoustic measurements for phonetic classification. In *Proc. ICSLP*, volume 1, Banff, Alberta, 1992.

[8] K. N. Stevens, S. Y. Manuel, S. Shattuck-Hufnagel, and S. Liu. Implementation of a model for lexical access based on features. In *Proc. ICSLP*, volume 1, pages 499–502, Banff, Alberta, 1992.

[9] Harvey M. Sussman, Helen A. McCaffrey, and Sandra A. Matthews. An investigation of locus equations as a source of relational invariance for stop place categorization. *JASA*, 90(3):1309–1325, September 1991.