Machine Learning Techniques for the Identification of Cues for Stop Place

Madelaine C. Plauché[*†] *and Kemal Sönmez*[†]

*University of California, Berkeley, USA

†SRI International, Menlo Park, USA

ABSTRACT

This paper is situated in a long line of phonetic studies that seek to determine and qualify the acoustic cues humans use to identify stop place. The present study draws from a database of 1500 CV tokens of American English and their values for the acoustic features thought to be cues for stop place identification, including (1) VOT, (2) energy of the burst and release, (3) spectrum at the burst, and (4) formant transitions into the following vowel. Decision trees are used to determine the relative invariance of these acoustic features, which indicates their potential to serve as useful cues for listeners cross-contextually. Decision trees thus allow the evaluation of vocalic effects on this hierarchy of features for the purpose of guiding classic perceptual confusion studies.

1. INTRODUCTION

50 years of phonetic research have explored the articulation of stops and their resulting acoustics. Acoustic information relevant to the identification of stop place resides in the relative amplitude[3,7] and spectral characteristics of the burst[1,3,10], as well as in formant transitions[5] and voice onset time before the following vowel[6]. Consonant confusion studies suggest that vocalic context is an important factor in the perception of neighboring stops. In the environment of a high front vowel (/i/), for example, cues such as formant transitions and VOT may be neutralized, resulting in a dependence on transient features that are secondary to the perception of stops in other contexts [8]. The present study estimates the effect of the vocalic context on the relative ranking of acoustic cues for stops in CV sequences using a machine learning algorithm that evaluates a large database of automatically extracted acoustic features.

2. DATABASE OF ACOUSTIC FEATURES

1500 CV tokens extracted from the careful speech of 7 American English speakers (4 men and 3 women) were collected. Each speaker was recorded in a sound booth uttering the frame sentence "Take __ for example," with s-initial words (Table 1) designed to elicit voiceless unaspirated stops [p], [t], [k] and cardinal vowels [i], [a], [u]. These recordings were digitized (sampling rate 16,000) directly into a Sun SparcC.

CV context	[i]	[a]	[u]
[p]	speak	spot	spook
[t]	steep	stop	stoop
[k]	skeet	scott	scoop

Table 1: Subject word list

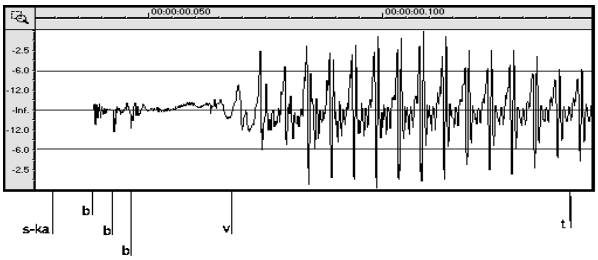

Figure 1: Hand-Labeled Waveform. This elicited [ka] token was hand-labeled for bursts, vowel onset, and final transitions.

Bursts (location and number), voicing onset, and transitions into the following vowel were hand-marked by a linguist (Figure 1) and the values of the following acoustic features were automatically extracted using ESPS software: (1) VOT, (2) relative energy and power of the burst (or bursts) with respect to the vowel, (3) linear and piece-wise linear fits to the spectrum at the burst, and (4) formant values and slopes of the following vowel. Trends from the resulting database are discussed in sections 2.1. to 2.4. by feature class.

2.1. Voice Onset Time (VOT)

Not only is VOT a known cue for the perception of manner (voiced vs. voiceless) in stops but, it also varies systematically within voiced and voiceless stops according to the place of articulation and the height of the following vowel[2,6].

In this study, voicing onset was defined as the first zero crossing of periodic noise (Figure 1). VOT for each CV token was extracted directly from the burst and voicing onset labels.

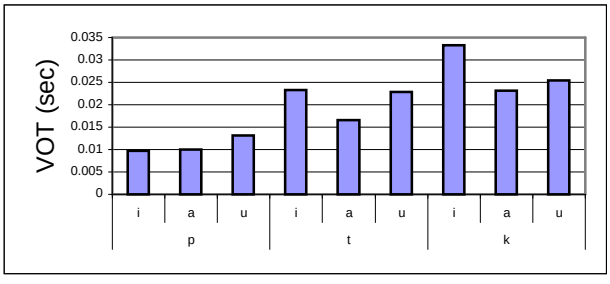

Figure 2: Average VOT for each CV context.

The average VOT for each stop place in this database follows trends found in previous VOT studies[6]. Voiceless labial stops have the smallest VOT, followed by alveolars, with velars

exhibiting the largest VOT. Stops of all places of articulation show higher VOTs on average preceding the high vowels [i] and [u], as the close oral constriction of high vowel offers greater impedance to the air escaping from the mouth, thus slowing the drop in oral pressure necessary for voicing [2]. VOT is predicted to be useful in a decision tree task of stop identification, both cross-vocalically and within a particular vowel context.

2.2. Energy of the Burst and Release

The amplitude of the burst and release, in particular, the amplitude of high frequencies, has been observed to cue listeners to place of articulation for voiceless stops. Ohde *et al*[7] found that lowering the relative amplitude of the high frequency component of a C[a] burst resulted in more [p] ratings than [t] ratings. Pattern playback studies[3] demonstrated that synthesized CV tokens with high frequency bursts were heard as /t/ across all vowel contexts, whereas bursts at lower frequencies were heard as /k/or /p/ depending on the relative position of the second formant of the vowel.

In order to capture the amplitude of each burst and release in the database, the absolute high frequency energy (above 3K) of the burst and release was normalized with respect to the maximum of the following vowel (Figure 3). As seen in section 2.1., the duration of the burst and release (VOT) varies across stop place, so the relative power (energy per second) of each burst across all frequency regions was also extracted (not shown here).

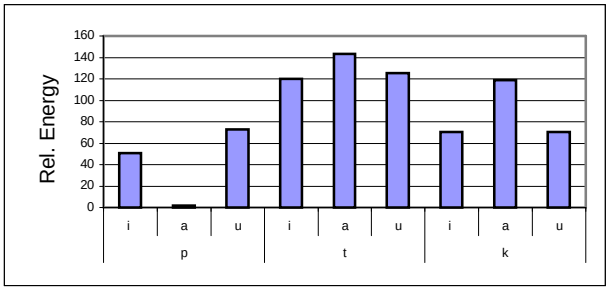

Figure 3: Relative High-Frequency Energy of Burst.

Labial stop bursts show the greatest contextual variation of relative high frequency energy. Labial stops preceding [a] have the smallest high frequency component at the burst. Preceding high vowels, however, the relative energy of the bursts are equivalent to those of voiceless velar stops, presumably due to the turbulence generated by the close constrictions of [i] and [u]. Alveolar stop bursts show the greatest relative high-frequency energy and power of all three places of articulation.

Another prominent feature of stop bursts, that may or may not be used by speakers to identify stop place, is the occasional presence of multiple bursts of voiceless velar stops, a result of the large area of the velar constriction and its relatively slow release[10]. The number of bursts per CV token was directly encoded in the database. However, the categorical aspect of multiple bursts is not necessarily a cue used by listeners to detect velarity. Indeed, given that multiple bursts are rare in spontaneous speech [9], it is more likely that listeners rely instead on the gross cues, energy and power, which capture the amount of energy generated by the release of a stop regardless of the actual number of bursts.

2.3. Spectrum at the Burst

The articulation of a voiceless stop and its release generate distinctive spectral characteristics at the burst that are somewhat invariant cross-contextually. Stop burst spectra for labials, alveolars, and velars have been described as “diffuse-rising” (majority of energy in the low frequency region), “diffuse-falling” (majority of energy in the high frequency region), and “compact” (peak of energy in the mid frequency region), respectively[1,10].

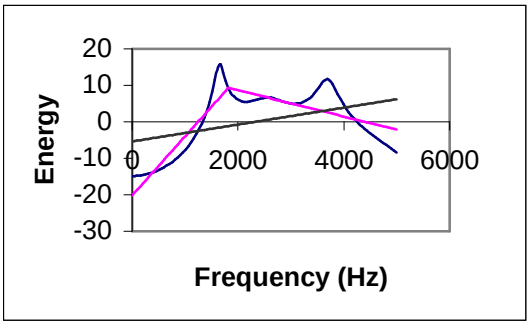

Figure 4: Burst Spectrum and Fitted Lines.

The spectrum of each CV burst (after pre-emphasis at 0.9), was fit to a linear (FIT 0) and a piecewise linear fit (FIT 1). Derived features include the slopes and y-intercepts 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 fits (see Figure 4 for feature extraction and Table 2 for database results).

Slopes and y-intercept values were highly variable across individual tokens, and so are expected to be useful to decision tree classification of stops only when combined with other spectral features.

Stop	Error Fit 0	Node (Hz)	Error Fit 1
[p]	1094	2600	412
[t]	1917	3441	414
[k]	2928	2583	986

Table 2: Average Values of Spectral Fit Features. Error Fit 0 and Error Fit 1 are the mean squared error of the linear fit and the piecewise linear fit, respectively.

The mid-compact peak of the velar yields high error rates for both linear fits and a mid-frequency node. Both labial and alveolar stops, characteristically described as diffuse-falling and diffuse-rising, yield lower error rates, but their average nodes indicate the frequency region with the greatest amount of energy at the burst: the low to mid region in the case of labials, and the mid to high region for alveolars. Whereas humans are thought to use burst spectra characteristics for stop place categorization only when more reliable cues such as formant transitions are obscured, the decision tree weighs all features equally. These spectral features should be just as useful for stop

place categorization as VOT, for example, due to their invariant properties.

2.4. Formant Transitions

Formant transitions, in particular F2 transitions, are considered primary cues for listener identification of stop place[3,4]. In conflicting cue tasks, listeners presented with an onset spectrum specifying one place spliced to formant transitions specifying another, rely primarily on the information provided by the formant transitions to determine stop place[5]. The closure of the lips for labial stops causes a lowering of all formants. Alveolars in CV contexts may show a small but rapid fall of formant transitions. During a velar release into a following vowel, the second formant and third formant are often close together, forming the characteristic "velar pinch". Additionally, the transitions from a velar are slower than corresponding labial and alveolar sounds, due to the mass of the articulator involved (i.e. the body of the tongue)[10].

Formant values (F1 through F4) for each CV token were extracted at 10 ms intervals from the onset of voicing of the vowel to 50 ms into the vowel using the ESPS formant tracker (Table 3). The formant tracker estimates speech formant trajectories by 10 ms frames, causing some errors at the onset of the voicing.

	p			**T**			**K**		
	i	**a**	**u**	**i**	**a**	**u**	**i**	**A**	**u**
F1	287	555	312	289	543	301	256	524	289
F2	2151	1065	1360	2271	1593	2033	2406	1629	1779
F3	2760	2449	2563	2847	2685	2837	3112	2550	2732

Table 3: Average Values for F1, F2, and F3 10 ms after voicing onset.

Though formant transitions have proven to be important cues for the detection of stop place by humans, problems with the formant tracker at the crucial region of voicing onset are expected to weaken the effects of this cue for classification by the decision tree.

3. DECISION TREE CLASSIFICATION

Classification over a set of stops is performed by feeding feature values from the database into a CART-style decision tree algorithm that repeatedly selects a single feature, which, according to an information-theoretic criterion (entropy), has the highest predictive value for detecting the correct stop place. For each classification task, the data was divided into a test and train set (roughly ¼ and ¾ of total data), with cross-validation set to 2, to eliminate overfitting on the relatively small set of data.

Section 3.1 contains the results of a decision tree classification of stops [p], [t], and [k]. The following sections (3.2. to 3.4) demonstrate shifts in relative ranking and accuracy of the extracted acoustic features for stop classifications depending on vocalic context.

3.1. Decision Trees for Stops [p], [t], and [k]

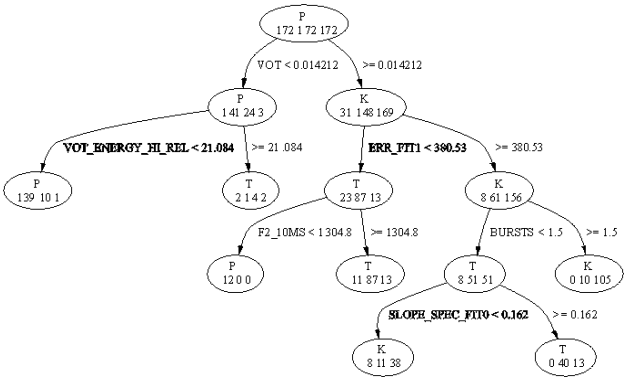

Figure 5: Decision Tree for [p], [t], [k] across all vowel contexts. Counts in each node are presented in the following order: p, t, k.

Figure 7 shows the classification of consonants [p], [t], and [k] for all three vocalic contexts. The overall classification accuracy for this tree is 84.3%.

As expected, VOT plays a major role in classifying stop place across vocalic contexts. Stops with short VOT's are successfully identified as [p]s, which are further separated from alveolars by the relatively high frequency energy feature. The error of the piecewise linear fit (ERR_FIT1) separates velars from other stops as predicted (Table 2). A low value for F2 onset picks out labials from other stops. Stops with multiple bursts (greater than 1.5) are identified as velars (and some alveolars). The slope of the linear fit to the spectrum grossly classifies alveolars (diffuse-rising) from velars.

3.2. Decision Trees for [a] Context

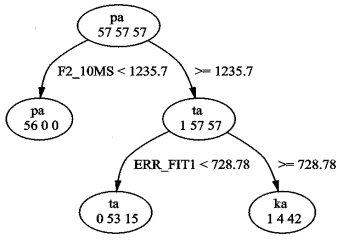

Figure 6: Decision Tree for [pa], [ta], [ka]. Counts are presented in the following order: pa, ta, ka. F2_10ms is the value of the F2 10 ms after the voicing onset of [a].

C[a] environments are thought to contain the most information about stop place, as stops in this context consistently show the lowest rates of confusion[8]. With only two features, F2 onset and error of the piecewise linear fit (ERR_FIT1), the decision tree classifies [pa], [ta], and [ka] with 88.3% accuracy (compare with 84% in 3.1.). Note that within this vowel context, despite problems with the ESPS formant tracker at voicing onset, F2 onset is the primary feature. This parallels its function as a primary cue in human perceptual studies[5].

3.3. Decision Trees for [i] Context

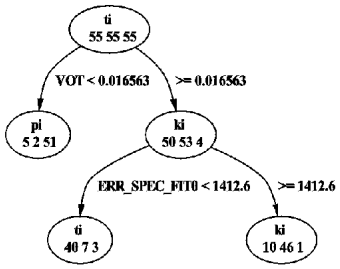

Figure 7: Decision Tree for [pi], [ti], [ki]. Counts are presented in the following order: ti, ki, pi.

As previously mentioned, the high front vowel [i] yields the highest confusion rates for listener-identification tasks, presumably due to the neutralization of primary cues such as formant transitions and the consequent reliance on transient features, such as spectral characteristics of the burst[8]. The decision tree classification of stops in this environment has the lowest accuracy rate for all contexts (83%). Much like the cross-vocalic case (section 3.1.), VOT is the most prominent feature, separating labial stops from alveolars and velars. The error of the linear fit to the burst spectrum is used to further separate the velars and alveolars. Formant features are not employed by the decision tree in this context.

3.4. Decision Trees for [u] Context

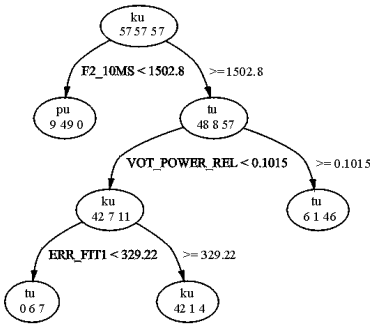

Figure 8: Decision Tree for [pu], [tu], [ku]. Counts are presented in the following order: ku, pu, tu.

The classification of stops in the [u] environment yields an accuracy rating in the same range as [a] (87.7%). As in the [a] context, the onset of F2 is a primary feature for distinguishing labials from other stops. Stops with high F2 onsets and high relative power of the burst and release are correctly classified as alveolar. Spectral errors further break down velars from other stops as predicted (Table 2).

Within the [u] context, as in the [a] context, F2 onset is the primary feature for tree classification, a result that mirrors proposed ranking of cues in human perceptual studies[5].

4. CONCLUSION

Although a direct comparison between tree and human discrimination can not be drawn, decision trees are an intuitive way to investigate what features are present in the signal and thus available to listeners in different contexts. The results presented above support an opportunistic model of perception, one in which listeners are thought to use whatever cues are available to them at the time. Further perceptual studies on the perceptual cues of human identification of stop place can rely on machine classification, such as decision trees, as an initial guide to areas of ambiguity in the acoustic signal responsible for human confusion errors. Additionally, relating statistical machine learning principles to existing phonetic research of inherently ambiguous speech signals can serve to improve aspects of front-end models of speech recognition.

Acknowledgments

This work was funded by the NSF Grant 9817243 and by the NSF-STIMULATE Grant IRI-9619921. The views herein are those of the authors and do not necessarily reflect those of the funding agencies.

5. REFERENCES

1. Blumstein, S., and Stevens, K., “Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants,” *JASA*, Vol. 66(4): 1001-1017, 1979.
2. Chang, S. S. “Vowel dependent VOT variation,” *Proc. of ICSLP 99*, San Francisco, 1999.
3. Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J. “Some experiments on the perception of synthetic speech sounds,” *JASA*, Vol. 24 (6): 597-606, 1952.
4. Delattre, P. C., Liberman, A. M., and Cooper, F. S. “Acoustic loci and transitional cues for consonants,” *JASA*, Vol. 27(4): 769-773, 1955.
5. Dorman, M. F. and Loizou, P. C. “Relative spectral change and formant transitions as cues to labial and alveolar place of articulation,” *JASA* 100(6): 3825-3830, 1996.
6. Klatt, D. H. “Voice onset time, frication, and aspiration in word-initial consonant clusters,” Journal of Speech and Hearing Research, 18:686-706, 1975.
7. Ohde, R. N. and Stevens, K. N. “Effect of burst amplitude on the perception of stop consonant place,” *JASA*, Vol. 74(3): 706-714, 1983.
8. Plauché, M. C., Delogu, C., and Ohala, J. “Asymmetries in Consonant Confusion,” *Proc. Of Eurospeech 97*, Rhodes, Vol. 4: 2187-2190, 1997.
9. Sönmez, K., Plauché, M. C., Shriberg, E. E., and Franco, H. “Consonant discrimination in elicited and spontaneous speech: A case for signal-adaptive front ends in ASR”, *Proc. of ICSLP*, Beijing, To appear.
10. Stevens, K. *Acoustic Phonetics*, Kluwer Academic Publishers, Boston, 1999.