CLASSIFICATION AND CLUSTERING OF STOP CONSONANTS VIA NONPARAMETRIC TRANSFORMATIONS AND WAVELETS

Basilis Gidas

Division of Applied Mathematics
Brown University
Providence, RI 02912

Alejandro Murua

Department of Statistics
University of Chicago
Chicago, Illinois 60637

ABSTRACT

We propose a new algorithmic method for the *classification* and *clustering* of the English six stop consonants /p, t, k, b, d, g/, on the basis of CV (Consonant-Vowel) or VC syllables data. The method explores two powerful tools: (1) a wavelet representation of the acoustic signal and its induced *"waveletogram"*, a time domain analogue of the spectrogram; (2) nonparametric transformations of the "waveletogram" and a nonlinear discriminant analysis based on these transformations. The procedure has yielded better rates of correct classification than previous methods. Moreover, it yields interesting two-dimensional *clustering plots* for stop consonants as well as for vowels. The clustering plots for vowels are as separating as those based on the first and second formants; we know of no other method in the literature that yields clustering plots for consonants.

1. INTRODUCTION

The Hidden Markov Models (HMM) approach, one of the most flexible and versatile frameworks, to speech recognition combines two basic models: the *acoustic model* and the *language model.* The approach has been highly successful in isolated-word speaker-dependent limited-vocabulary tasks. Recent systems have addressed these restrictions with various degrees of success, but there remain fundamental difficulties in dealing with more "natural" speech. In the view of some researchers (including ourselves), current implementations of HMM have at least two weaknesses: (a) one at the "low-level", i.e. how the data interact with the phoneme models (as well as the nature of the phoneme models); (b) and one at the "high-level", i.e. the use of artificial grammars.

This work was partly supported by ARO Grant DAAH04-93-G-0046, ONR Grant N00014-91-J-1021, and ARPA via ARL MDA972-93-1-0012

In [1, 2] we introduce a framework based on modern *nonparametric* techniques for estimation and prediction, to deal with the "low-level" difficulties. The procedure is a *time domain* approach and leads to a new *acoustic model,* i.e. to a new type of phoneme models and a new relation between these models and the acoustic data.

The framework is motivated by both the acknowledgement that the acoustic signal contains *non-stationary* and *nonlinear* segments, and the fact that nonlinearities contain information that is important for recognition. Most speech recognizers, including HMM, assume that short time segments are stationary and linear. Hence, they are analyzed via Fourier Transform (FT) methods and linear models (such as Linear Predictive Coding, LPC). These procedures are adequate in some parts of the signal (e.g. steady state of vowels), but not in others: non-stationarities in transition regions make FT questionable; and linear models cannot capture useful information contained in nonlinearities.

In order to isolate the effects of non-stationarities and nonlinearities from other effects of the signal, we consider the problem of classifying (recognizing) the six stop consonants /p, t, k, b, d, g/ and vowels on the basis of CV or VC syllables. The problem has the additional advantage of not requiring a language model, and hence it is a good case for testing low-level procedures. In addition to being directly relevant to speech recognition, this problem has important implications [3] in phonetic theories, theories of perception, developmental psychology and auditory physiology.

The stop consonants problem has been studied with a variety of methods [4, 5, 6, 7], including neural networks [4, 8]. Its central difficulty lies in the non-stationary and nonlinear structure of the signal in the transition and burst regions. Our handling of these effects explores two powerful tools: (1) a wavelet representation of the acoustic signal and its induced *"wave-*

0-7803-2431-5/95 $4.00 © 1995 IEEE

letogram", a time domain analogue of the spectrogram; and (2) nonparametric transformations of the waveletogram, and a nonlinear discriminant analysis based on these transformations.

Our procedure has yielded (see Section 4) better correct classification rates than previous methods [4, 5, 8] for stop consonants and vowels. Further, previous methods have focused primarily in the discrimination within voiced /b, d, g/ or unvoiced /p, t, k/ stop consonants only. Our numerical results are in complete agreement with psychophysical experiments [9, pp. 190–191], which indicate that adults and infants discriminate correctly the three voiced stop consonants and the vowels when listening to about 100 ms of speech. An important additional aspect of our procedure is that it yields interesting two-dimensional *clustering plots* for vowels as well as for stop consonants. The clustering plots for vowels are comparable to those based on the first and second formants. The clustering plots for consonants separate the stop consonants into clusters according to the two *phonetic features* that differentiate them: the *place of articulation* and the phonetic feature *voicing*, which is usually characterized by a single parameter known as *Voice Onset Time* or VOT.

Next, we describe the basic components of our classification procedure, which will be described in Section 3. Section 4 reports the results of two experiments designed to test the effect of multiple speakers and different vowel environments in the classification of stop consonants based on our procedure.

2. THE CLASSIFICATION ALGORITHM

2.1. The Wavelet Representation and the Waveletogram

Let $X = \{X_t : t = 0, 1, \ldots, 2^N - 1\}$ be the sampled speech signal in the time domain. Finite Discrete Wavelet Transforms (FDWT) are indexed by two integers R ($2R+1$ is the support width) and j_0 (low resolution cut-off); and are either "periodic" or "boundary adjusted" [10, 11]. For a periodic FDWT, X is represented by

$$X_t = Z_{-1,0} W_{-1,0}(t) + \sum_{j=0}^{N-1} \sum_{k=0}^{2^j-1} Z_{j,k} W_{j,k}(t) \qquad (1)$$

where $Z_{-1,0}$, $Z_{j,k} : j = 0, \ldots, N-1, \ k = 0, \ldots, 2^j - 1$, are the wavelet coefficients of X. The basis elements $W_{j,k}(t)$ are related to Daubechies' compactly supported wavelets [10]. For $j_0 < j << N$, $0 << k << 2^j$, $W_{j,k}(t)$ is concentrated at *scale* 2^j and *location* $t = k2^{N-j}$. For other properties of $W_{j,k}(t)$ see [10, 11].

The *"waveletogram"* corresponding to (1) is simply the plot of $\{|Z_{j,k}|\}$ in the (j,k)-plane. After properly smoothing the waveletogram, we extract a small number of parameters $Z = (Z_1, Z_2, \ldots, Z_T)$; in our experiments $T = 21$. These are the parameters (data) used in our classification and clustering algorithm.

2.2. Nonparametric Transformations

Let M be the number of units to be classified (for the stop consonants $M = 6$). We construct: (1) a sequence of parameters $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{M-1} \geq \lambda_M = 0$, with $\lambda_1 < 1$, (2) a matrix of constants

$$g_n(m), \quad m = 1, \ldots, M, \quad n = 1, \ldots, M-1$$

and (3) a matrix of nonlinear functions

$$f_{i,n}(Z_i), \quad i = 1, \ldots, T, \quad n = 1, \ldots, M-1.$$

These are solutions of an eigenvalue problem [1, 12] whose formulation involves a smoothing procedure. In our experiments we used the well-known *super-smoother*. Our classification and clustering [1, 13] are based on

$$E_n(Z) = \sqrt{\lambda_n/(1-\lambda_n)} \sum_{i=1}^{T} f_{i,n}(Z_i)$$

$$C_n(m) = \sqrt{1/(1-\lambda_n)}\, g_n(m), \quad m = 1, \ldots, M$$

for $n = 1, \ldots, M-1$.

2.3. Classification and Clustering Plots

Let

$$D(Z; m) = \sum_{n=1}^{M-1} [C_n(m) - E_n(Z)]^2$$

The *classification* rule is: given $Z = (Z_1, \ldots, Z_T)$, choose $m \in \{1, 2, \ldots, M\}$ so that it minimizes $D(Z; m)$.

From the point of view of this classification rule, the high-dimensional vector $Z \in \mathbb{R}^T$ has been replaced by the low-dimensional vector

$$(E_1(Z), E_2(Z), \ldots, E_{M-1}(Z)) \in \mathbb{R}^{M-1} \qquad (2)$$

Thus, the space $\mathbb{R}^{M-1}$ may be interpreted as the *hypotheses space* (or *class space*) in which the vector

$$(C_1(m), C_2(m), \ldots, C_{M-1}(m)) \in \mathbb{R}^{M-1}$$

determines the *center of hypothesis* m, and vector (2) for various realizations of Z (from the conditional probability $P(Z|m)$), defines the *"cloud"* or *"scatterplot"* of hypothesis m.

The map that to Z corresponds vector (2) may be viewed as a *nonlinear dimensionality reduction map.* The success of clustering depends, naturally, on how well the hypotheses (i.e. their scatterplots) are *separated* in the hypotheses space $\mathbb{R}^{M-1}$. In addition, separation of hypotheses should also be "visible" in lower-dimensional subspaces of $\mathbb{R}^{M-1}$.

In principle, one could use *Projection Pursuit* (or some other dimensionality reduction algorithm) to find "optimal" directions and subspaces in $\mathbb{R}^{M-1}$. We have not explored this possibility, but we have investigated clustering properties in subspaces defined by choosing a subset of the coordinates of vector (2). More precisely, let $r < M-1$, and fix $n_1 < n_2 < \cdots < n_r$, where $n_1, \ldots, n_r \in \{1, \ldots, M-1\}$. The subspace

$$(E_{n_1}(Z), \ldots, E_{n_r}(Z)) \in \mathbb{R}^r$$

may be viewed as an *r-dimensional hypotheses space.* The center of hypothesis m in this space is then

$$(C_{n_1}(m), \ldots, C_{n_r}(m)) \in \mathbb{R}^r$$

For $r = 2$, the two-dimensional *clustering plots* are the scatterplots in the $(E_n, E_{n'})$, $n' \neq n$, planes. Typically, the (E_1, E_2)-plane provides the best separation. In our experiments of Section 4 we obtain interesting clusterings of the stop consonants into *voiced* and *voiceless,* and into *labial, alveolar* and *velar* consonants.

3. CLASSIFICATION OF SPEECH SOUNDS

In this section we describe how the above nonlinear classification rule is applied to speech signals.

We divide the acoustic signal in L consecutive overlapping frames $\mathcal{L} = \{1, 2, \ldots, L\}$. In each frame $\ell \in \mathcal{L}$ we extract T relevant variables $Z^{(\ell)} = (Z_1^{(\ell)}, \ldots, Z_T^{(\ell)})$, $\ell = 1, \ldots, L$, as explained in Subsection 2.1. The entire signal is then replaced by $Z = \left\{Z^{(\ell)} : \ell = 1, \ldots, L\right\}$. In every frame $\ell \in \mathcal{L}$, the functions

$$\left\{g_n^{(\ell)}(m) : m = 1, \ldots, M, \;\; n = 1, \ldots, M-1\right\},$$

$$\left\{f_{i,n}^{(\ell)}(\cdot) : i = 1, \ldots, T \;\; n = 1, \ldots, M-1\right\},$$

are estimated and used to form the nonlinear *discriminant measure*

$$D^{(\ell)}(Z^{(\ell)}; m) = D(Z^{(\ell)}; m)$$

For a total number L of frames, we employ the *overall discriminant measure*

$$D_L(Z; m) = \sum_{\ell=1}^{L} D^{(\ell)}(Z^{(\ell)}; m). \qquad (3)$$

The classification rule is then: given $Z = \{Z^{(\ell)} : \ell = 1, \ldots, L\}$, choose the hypothesis m that maximizes (3). If more than one hypothesis attains the minimum, then choose randomly among the competing hypotheses, with equal probability.

4. EXPERIMENTS

We describe two experiments using two speech data sets. The first experiment was designed to test the effect of multiple speakers (three female ant two male speakers) in the classification of the six stop consonants in a fixed vowel environment. The second experiment was designed to test the effect of different vowel contexts (five vowels) in the classification of stop consonants within a single (male) speaker.

4.1. Experiment 1

We used 395 CV tokens (uttered by 3 female and 2 male speakers) of the six stop consonants /p, t, k, b, d, g/ followed by the vowel /a/. 252 CV tokens were used for training, and 143 for testing. The average correct classification rate was over 96% when 50 to 60 ms of the acoustic signal (from the beginning of the burst) were used. This time scales are consistent with the empirical values of the VOT. The number of tokens per speaker were unequal; in fact, two speakers had a small number of tokens. In this sense, the result is nearly speaker independent.

4.2. Experiment 2

In this experiment, we used 424 CV tokens (uttered by one male speaker) of the six stop consonants /p, t, k, b, d, g/ followed by the vowels /a, e, i, o, u/, i.e. a total of 30 CV categories. 282 CV tokens were used for training, and 144 for testing. The average correct classification rate for stop consonants was over 93% when more than 60 ms of the acoustic signal were used, and over 95% when 90 to 120 ms were used (the maximum time considered in our experiments was 220 ms from the beginning of the burst). For the vowels, the average correct classification rate improves steadily with the total number L of frames considered in the algorithm, achieving average error rates less than 1% when 200 ms of the speech signal were used. The tradeoff to obtain good classification rates for both the stop consonants and the vowels appears to be when 190 to 210 ms of the acoustic signal were used, with average correct syllable classification rate over 94%.

4.3. Clustering

We have used the clustering functions $E_n(\cdot)$ defined in Section 2 to construct two-dimensional clustering plots for the vowels as well as for the stop consonants. The clustering of the five vowels /a, e, i, o, u/ in the two-dimensional vowel space (E_1, E_2) is very reminiscent of the clustering plots of vowels [9, p. 178] based on the first and second formants. This indicates a close relation between the first two clustering functions E_1, E_2, and the first two formants. The clustering of the six stop consonants /p, t, k, b, d, g/ in the two-dimensional consonant space (E_1, E_2) shows simultaneously a separation between the *voiced* stop consonants /b, d, g/ and *voiceless* stop consonants /p, t, k/, and a division among the *labial* consonants /p, b/, *alveolar* consonants /t, d/ and *velar* consonants /k, g/, i.e. a division according to *place of articulation*. In this sense, the first two functions E_1, E_2, represent acoustic correlates corresponding to the phonetic features voicing and place of articulation. We have not found in the literature any other acoustic functions directly derivable from the acoustic signal that may be used to obtain two-dimensional clustering plots for the stop consonants.

5. REFERENCES

[1] A. Murua. *Optimal Transformations for Prediction in Continuous Time Weakly Stationary Processes and Applications to Phoneme Recognition.* PhD thesis, Division of Applied Mathematics at Brown University, Providence, Rhode Island, 1994.

[2] B. Gidas and A. Murua. Nonlinear transformations for prediction and applications to speech recognition. In preparation.

[3] S. E. Blumstein. Towards a theory of acoustic invariance in speech, 1983. Lecture Notes, Brown University.

[4] J. A. Anderson, M. L. Rossen, S. R. Viscuso, and M. E. Sereno. Experiments with representation in neural networks: Object motion, speech, and arithmetic. In H. Haken and M. Stadler, editors, *Synergetics of Cognition.* Springer-Verlag Berlin, 1990.

[5] C. L. Searle, J. Z. Jacobson, and S. G. Rayment. Stop consonant discrimination based on human audition. *Journal of the Acoustical Society of America*, 65:799–809, 1979.

[6] K. S. Nathan and H. F. Silverman. Time-varying feature selection and classification of unvoiced stop consonants. *IEEE Trans. Speech Audio Process.*, 2(3):395–405, 1994.

[7] K. N. Stevens and S. E. Blumstein. Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 64:1358–1368, 1978.

[8] M. L. Rossen, L. T. Niles, G. N. Tajchman, M. A. Bush, J. A. Anderson, and S. E. Blumstein. A connectioninst model for consonant-vowel syllable recognition. In *ICASSP-88*, pages 59–66, 1988.

[9] P. Lieberman and S. Blumstein. *Speech Physiology, Speech Perception, and Acoustic Phonetics.* Cambridge University Press, 1988.

[10] I. Daubechies. *Ten Lectures on Wavelets.* SIAM, 1992.

[11] A. Cohen, I. Daubechies, B. Jawerth, and P. Vial. Multiresolution analysis, wavelets and fast algorithms on an interval. *Comptes Rendus Acad. Sci. Paris, Série I*, 316:417–421, 1993.

[12] L. Breiman and R. Ihaka. Nonlinear discriminant analysis via scaling and ACE. Technical report, University of California, Berkeley, Dept. of Statistics, 1984.

[13] B. Gidas and A. Murua. Classification and clustering of speech units via nonparametric transformations, 1995. To appear in IMA, University of Minnesota.