CONTEXT DEPENDENT RECOGNITION OF SPANISH STOPS

HORACIO FRANCO

JORGE A. GURLEKIAN

Laboratorio de Investigaciones Sensoriales. CONICET.
CC 53 (1453) Buenos Aires, Argentina.

ABSTRACT

The recognition performance using either the entire or selected portions of running spectra obtained from intervocalic Spanish stops is presented in the framework of a recognition system based on segmentation and context dependent statistical classification.

In order to capture the dynamic nature of this context dependent sounds, critical band spectra were uniformly grouped in a temporal sequence of contiguous segments.

A set of conditional probability density functions were estimated for the spectra belonging to each segment for every different vocalic context.

The contribution to the recognition performance of different segments was evaluated and related to the relevant acoustic features.

The performance of the recognizer was also studied under different degrees of context dependence.

Results for unvoiced and voiced stops were obtained in a speaker dependent manner using a data base consisting of 2592 emissions of the stops / p, t, k, b, d, g/, embedded in VCVCVCV sequences where the V's were the vowels /a, i, u/ uttered by two male Argentine speakers.

1. INTRODUCTION

Among the stop consonant recognition systems the best performance at present was obtained in those works which used contextual information /1/ or sequences of short time spectra as features for recognition /2/.

In this work we combined these two characteristics in the framework of a statistical approach. Our basic objective was to evaluate the recognition performance of an automatic recognition system for the Argentine Spanish stops uttered in intervocalic position based on the use of a context-dependent statistical classifier, and using the whole set of spectra along the acoustical realization as features for the classification.

With this objective we also evaluated the contribution of portions of the spectral sequence for the recognition of the stops under different degrees of context dependence, i.e., either the following or preceeding or both vowel classes which act as the a priori information for the classifier.

2. ACOUSTICAL DESCRIPTION

In the case of voiced stops, the intervocalic realization is an approximant as was defined by Ladefoged /4/ with no signs of burst and with a week evidence of consonant release. Moreover, the stops appear as a slight acoustic variation of the transition between two particular vowels. The acoustic correlates of the coarticulation of the VCV sequences follow aproximately the typical formant patterns obtained by Ohman /3/ for Swedish and American English. Quasi-stationary portions of vowels are slightly affected by the consonant and transconsonantal vowel, while the formant patterns of the occlusive portions along the energy dip are dependent of the two, initial and final, vowel classes.

On the other hand, unvoiced stops show a burst following the silent gap but weaker than in American English emissions.

3. SPEECH DATA AND SPEECH ANALYSIS

The speech data base consisted of 2592 emissions of the intervocalic stops /p, t, k, b, d, g/ combined with the vowels /a, i, u/ in all combinations, i.e., 9 vocalic contexts for the VCV sequences. The speech data was uttered by two male Argentine speakers.

As a step to test the system with running speech, three different consonants sharing the context vowels were produced embedded in VCVCVCV nonsense utterances. In this way, continuous utterances of 800 msec (on the average) were produced and the stress pattern of the VCV sequences had a greater degree of variation.

The speech samples were recorded in a low noise environment, low pass filtered to 5 kHz, sampled at 10 kHz and digitized with a 12 bit A/D converter. Each utterance was preceeded and followed by short segments of background noise and stored in separate waveform files. From the speech waveform stored on disk the following parameters were extracted every 10 msec through a sliding 25.6 msec Hamming window of the preemphasized waveform.

a) Short time energy expressed in dB.

b) Critical Band Spectra. From a 128 point short-time DFT spectra (40 Hz resolution), the energy output of an auditory filter bank was obtained following the method described by Moore /5/ but using the Zwicker's /6/ critical bandwidths and critical band rate scales instead (Fig. 1). Energy at each critical filter was obtained as a weighted sum of DFT energy values. The energy transfer function for each filter was the rounded exponential:

$$W(g) = (1 + g p) \, \mathrm{Exp}(- g p)$$

were $g = \mathrm{abs}(f - fc) / fc$ and $p = 4 fc / BW(fc)$ with BW(fc) the bandwidth of the critical filter centered at fc. Center frequencies were spaced one critical bandwidth starting from 100 Hz till 4500 Hz, in this way 18 filters resulted. These spectra were expressed in dB and normalized by substracting the linear mean of each in logarithmic scale.

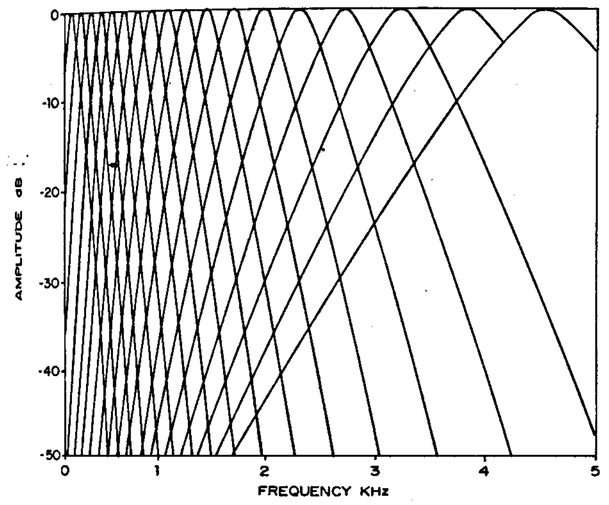


Fig. 1. Responses of the 18 individual channels of the critical filter bank.

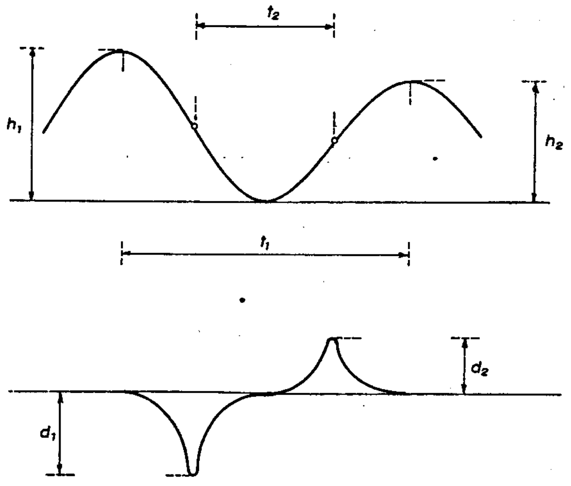


Fig. 2. On top: log-energy contour showing the measurements used for the definition of dip features (Fn); F1 = h1+h2; F2 = t1, and F3 = t2. Bottom: First derivative contour of the log-energy showing one additional feature: F4 = d1+d2.

4. SEGMENTATION

The proposed scheme was: first, to find reliable points for the temporal location of the consonant using the short time energy contour, and secondly to characterize its dynamic nature through different probability distributions associated with portions of the spectral sequence.

Previous research /7/ showed the relevance of the temporal amplitude contour dips for the perceptual detection of the voiced stops, so, the segmentation strategy was based on the use of this acoustical event.

The segmentation was accomplished as follows: first, a dip detection was performed finding the local maxima and minima in the log-energy contour wich was smoothed via six passes through a zero phase digital filter with a three point triangular impulse response. With this degree of smoothing a great percentage of spurious peaks and dips were removed although the peaks and valleys associated with vowels and this kind of consonants were preserved. Following this, a dip classification was performed. For each dip and associated left and right peaks a set of features characterizing its width, depth, and abruptness were measured from the log-energy contour and its unsmoothed derivative (Fig. 2).

A statistical classifier /8/ assuming a Gaussian multivariate probability distribution was trained using each half of the data, then the dips from the other half were classified in the following categories: unvoiced dip, voiced dip, and dip non valid.

The consonantal portions were defined in valid peak-dip-peak sequences, for each of them, the points of maximum slope at the consonant closure and release were located in the log-energy contour.

In order to characterize the time varying spectral pattern of the consonant, the time scale was linearly mapped to seven segments and the spectra belonging to each segment were associated with a single probability distribution. The two points of maximum slope were used as anchor points of the seven linearly distributed segments with the second and sixth segment tied to those anchor points. Five segments span between the anchor points including them, and two others were extrapolated at the extremes.

Even though the unvoiced stops in Argentine Spanish present a clear release, the same scheme as for the voiced stops was used, given that in VCV contexts, the spectra belonging to both the VC transition and the CV transition could provide evidence for the recognition.

This procedure provided an aproximate time normalization for the dynamic portions and a simple method of alignment for the test and reference patterns.

Vocalic segments were defined either between two consecutive consonantal portions or between the endpoints of the utterance and the closest consonantal portion.

5. THE CONSONANT CLASSIFIER.

The features selected for training and classification were normalized critical band spectra obtained at the seven segments located along the log-energy dip as defined above. Given the unstationary and context-dependent nature of the consonantal realizations, different probability distributions were assumed for the spectra at each one of the seven segments, for each one of the different vowel contexts considered.

A bayesian context-dependent classifier /8/ was designed according to the following assumptions:

a) the class conditional probability density functions of the spectra at each segment are Gaussian independent,

b) there is statistical independence between spectra,

c) the vowel and consonant classes are equiprobable and independent.

In the training phase the speech data from each speaker were split in two halves. From each half, maximun likelihood estimates of the mean vector and the assumed diagonal covariance matrix of each class conditional probability density function were obtained after a supervised segmentation.

In the recognition phase the classifier was run over each data half with the parameters obtained from the other half. Under the assumption that the classes of adjacent vowels are known, the parameters of the consonant classifier were chosen among those obtained in the training phase for the different vowel contexts considered. To accomplish the recognition, a conditional log-likelihood Lj/c for each consonant class j and the given context c, was obtained as:

$$L_{j/c} = - \sum_{k} \sum_{i=1}^{18} \left[\frac{(Y_{ik} - M_{isjc})^2}{V_{isjc}} + \log(V_{isjc}) \right]$$

with $s = S(k)$

where Yik is the value of the normalized spectra corresponding to the ith critical band at time index k, and Misjc and Visjc are respectively the mean and variance associated with the ith critical band of the spectra corresponding to the segment s for the jth consonant class in the assumed known context c. The segment index s is obtained from the time index k through the segmentation mapping s = S(k).

So, L is like a "global distance" computed as the sum over all the segments, of the weighted euclidean distances between the spectra belonging to each segment and the corresponding mean spectra, plus a segment dependent term. The weights are the inverses of the variances of the corresponding spectral samples.

The classifier performed three-way /p, t, k/ or /b, d, g/ discriminations. The recognized consonant corresponded to the largest likelihood.

The classifier was run using the spectra belonging to all the segments and alternatively using the spectra obtained from single segments to evaluate their particular recognition performance.

The degree of context dependence was given by the classifier training and operation according to the consideration of three alternative cases. In the first, the information of the preceeding and following vowel classes is used to select different probability distributions. This case will be referred to as VCV recognition. In the second and third cases the information of only the preceeding or the following vowel classes is used. These cases will be referred to respectively as VC or CV recognition.

The context vowels were recognized, using the spectra belonging to the vocalic segments, by means of a similar statistical classifier that uses a single probability distribution for each vowel class given its assumed quasi-stationarity and context independent nature.

6. RESULTS

6.1 Segmentation.

The performance of the dip detector and classifier to discriminate between valid consonant dips or invalid dips was of 99.7% for speaker 1 and 98.7% for speaker 2. The voiced-unvoiced discrimination among the valid consonant dips detected, reached a 96,1% for speaker 1 and 97.5% for speaker 2.

6.2 Consonant Recognition.

The recognition rate of the VCV, CV and VC cases using the spectra corresponding to all and single segments are presented for each speaker and for the unvoiced and voiced stops in Table I and Figs. 1 to 4.

	VCV	CV	VC
Sp.1			
Unvoiced stops	92.7	90.4	80.3
Voiced stops	88.5	76.4	73.6
Sp.2			
Unvoiced stops	94.3	94.2	88.3
Voiced stops	90.8	75.5	77.0

Table I. Recognition scores (%) corresponding to the performance when the information of all segments is used.

Under the VCV context condition an evaluation of the individual spectral segments was obtained. As it can be seen the recognition performance is relatively uniform over the seven segments in the case of voiced sounds. For the unvoiced stops it is clear the highest contribution showed by the segments corresponding to the stop release and the transitional part towards the following vowel.

When the knowledge of the context that was accounted for the classifier training and operation was restricted to only the following vowel (CV recognition), the voiced stops aproximately doubled the error rate obtained for the VCV case. However for the unvoiced stops the performance held similar values.

On the other hand for the VC recognition, the voiced stops gave similar results as for the CV recognition showing that in this case there is no clear preference for the accounting of the following or previous vowel context. This was not the case for the unvoiced stops. Accounting of the following vowel clearly gave a better performance than using the preceeding vowel.

With reference to the performance of the individual spectral segments under the different degrees of context dependence, it can be observed that the segments close to the vowel not considered in the VC and CV context conditions significatively lowered their recognition scores, while the segments located near the opposite vowel approached the values corresponding to the VCV case.

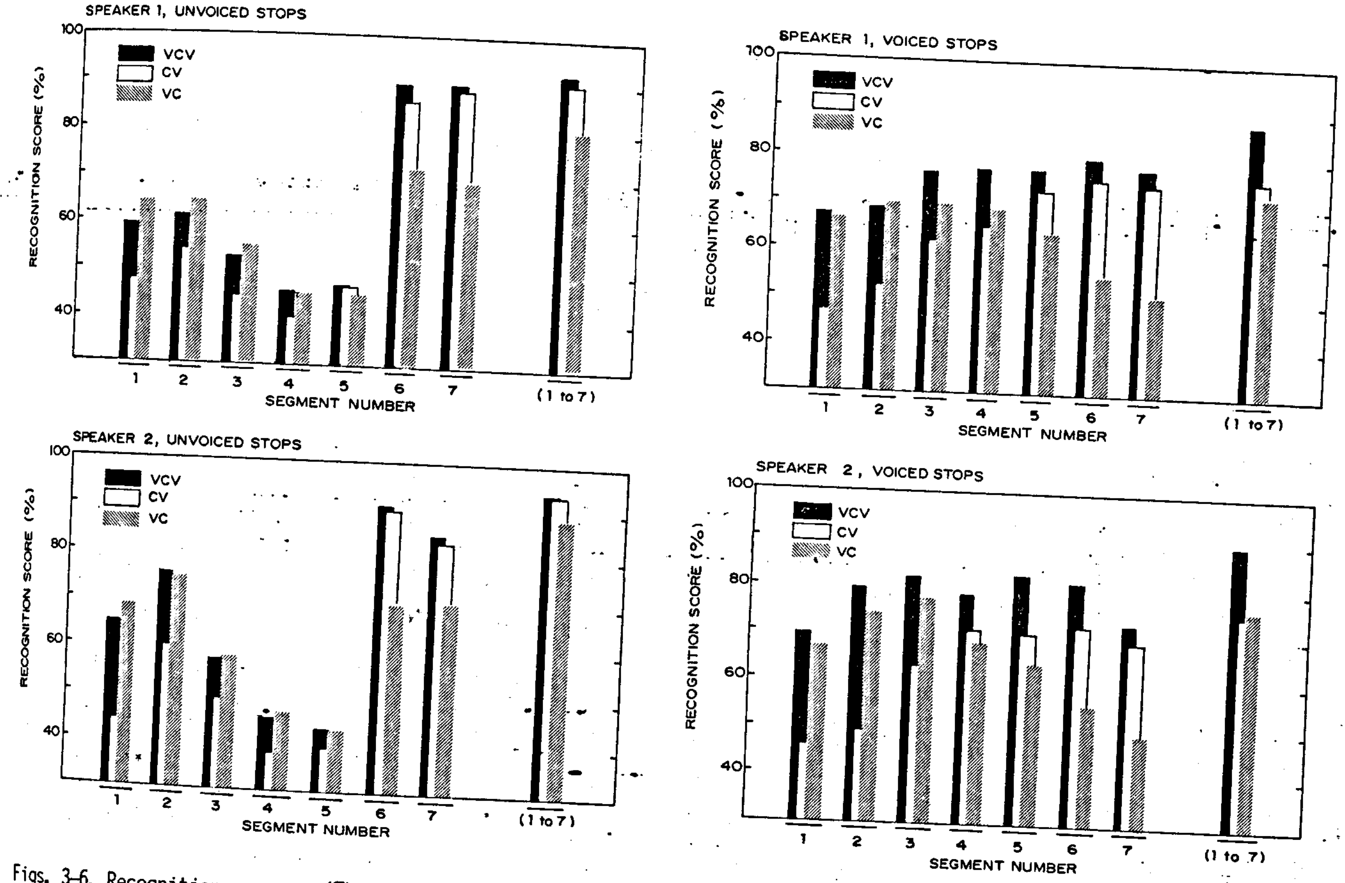
The performance results discriminated for every vocalic context for the VCV recognizer ordered by decreasing performance are presented in Table II. The average recognition rate ranged from 99.6% for the /a-a/ context to 81.8% for the /i-u/ context.

	a-a	a-i	i-a	u-a	i-i	a-u	u-u	u-i	i-u
Sp.1									
U	100	98.6	91.7	98.6	94.4	95.8	81.9	81.9	91.7
V	100	90.2	100	88.9	94.4	84.7	76.4	77.8	83.3
Sp.2									
U	100	98.6	93.1	93.1	93.1	95.8	94.4	93.1	87.5
V	98.6	95.8	97.2	98.6	87.5	91.7	95.8	86.1	65.3

Table II. Recognition scores (%) for unvoiced (U) and voiced (V) stops discriminated by vocalic contexts.

Considering that the contribution of segments corresponding to the silent gap in the unvoiced sounds could introduce noisy information to the classifier the recognizer was also run using only selected segments such as, numbers 2, 6, and 7 which presented the best individual scores. For this case the recognition scores are presented in Table III.

Given the limited data available to train the recognizer the supression of the noisy information effectively improved the scores for unvoiced stops.



Figs. 3-6. Recognition scores (%) using the spectra from single and all segments (1-7) under different degrees of context dependence (VCV, CV, and VC), presented for both speakers.

The recognition of the context vowel classes was accomplished with a high performance giving only one error over the total of 3456 vowels recognized for both speakers.

(VCV)	Speaker 1	Speaker 2
Unvoiced stops	95.1	96.5

Table III. Recognition scores using selected segments.

8. CONCLUSION

In this work we have tested a statistical approach to the recognition of unvoiced and voiced stops.

The best results for the voiced stops were obtained when the "two vowel" context dependent classifier was used. For the unvoiced stops the classifier achieved similar higher scores when using either the "two vowel" context or only the following vowel context.

Regarding the performance of the individual segments of the spectral sequence, the results obtained under the VCV vowel context condition showed that for the voiced stops the information was quite uniformly distributed along the whole VCV pattern. For the unvoiced stops there was a clear dominance of segments around the burst and transitions to the following vowel.

These results suggest that at least for Spanish, an intervocalic consonant recognition strategy could be based in the use of CV units for unvoiced stops but it should be based on larger units as the VCV for the voiced stops.

The authors wish to thank Dr. B. Cernuschi for his helpful suggestions.

10. REFERENCES

/1/ P. Demichelis, R. De Mori, P. Laface and M. O'Kane, "Computer Recognition of Plosive Sounds Using Contextual Information", IEEE Trans. Acoust., Speech and Signal Processing, Vol ASSP-31, 359-377, 1983.
/2/ G. E. Kopec, "Voiceless Stop Consonant Identification Using LPC Spectra", IEEE Trans. Acoust. Speech and Signal Processing, ICASSP84. March 19-21, 1984.
/3/ S. E. G. Ohman, "Coarticulation in VCV Utterances: Spectrographic measurements", J. Acoust. Soc. Am., 39, 151-168, 1966.
/4/ P. Ladefoged, "A Course in Phonetics", Nueva York, Harcout Brace Jovanovich, Inc., 1975.
/5/ B. Moore, B. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns", J. Acoust. Soc. Am., 74, 750-753, 1983.
/6/ E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency", J. Acoust. Soc. Am.,68, 1523-1525, 1980.
/7/ H. Franco, J. A. Gurlekian, "Recognition of Spanish intervocalic consonants", J. Acoust. Soc. Am., Vol. 77 S1, S27, 1985.
/8/ R. Duda, P. Hart, "Pattern Classification and Scene Analysis", Wiley-Interscience, 1973.