# RECOGNITION OF UNVOICED STOPS FROM THEIR TIME-FREQUENCY REPRESENTATION

*Maria Rangoussi and Anastasios Delopoulos*

Computer Science Division,
Department of Electrical Engineering,
National Technical University of Athens,
Athens GR-15780, GREECE
e-mail: maria@ntua.gr

## ABSTRACT

Recognition of the unvoiced stop sounds /k/, /p/ and /t/ in a speech signal is an interesting problem, due to the irregular, aperiodic, nonstationary nature of the corresponding signals. Their spotting is much easier, however, thanks to the characteristic silence interval they include. Classification of these three phonemes is therefore proposed in the present paper, based on patterns extracted from their time - frequency representation. This is possible because the different articulation points of /k/, /p/ and /t/ are reflected into distinct patterns of evolution of their spectral contents with time. These patterns can be obtained by suitable time - frequency analysis, and then used for classification. The Wigner distribution of the unvoiced stop signals, appropriately smoothed and subsampled, is proposed here as the basic classification pattern. Finally, for the classification step, the Learning Vector Quantization (LVQ) classifier of Kohonen is employed on a set of unvoiced stop signals extracted from the TIMIT speech database, with encouraging results under context- and speaker- independent testing conditions.

## 1. INTRODUCTION

The quality of recognition of basic acoustic-phonetic units is important to the success of the higher levels in the speech recognition task. Composite continuous recognition systems such as SPHINX, [7], use features carrying spectral information, like LPC or cepstral coefficients, for their baseline subsystems. Voiced phonemes are thus successfully recognized due to their almost periodic nature, while recognition of unvoiced phonemes has been far more difficult up to now, [3]. Methods that exploit contextual information from the neighboring phonemes have been resorted to, as a consequence.

In the present paper we deal with the analysis and classification of the sounds /k/,/p/ and /t/, termed *unvoiced plosives* or *unvoiced stops*, [9]. These sounds include an initial silence part, corresponding to a closure of the vocal tract, during which vocal cords do not vibrate, followed by an explosion part, corresponding to a bursty air release from the lungs, when the vocal tract suddenly opens. The different closure points, velum, lips or teeth, produce /k/,/p/ and /t/, respectively. Unvoiced stops, therefore, correspond to irregular, nonstationary signals, with a typical silence part and a random explosion part.

The silence part allows for a relatively easy spotting of the unvoiced stops within a speech signal. Results in this direction, employing high-order statistics of speech, are reported in [10]. The explosion part, however, is far more informative with respect to the classification of the three sounds after spotting them. The random and nonstationary nature of this part, along with its short duration in time (50-100 msec typically), prompts the use of time - frequency representations for their analysis and for feature extraction in view of classification. Previous work focuses on a specific context in a vowel-consonant combination, and resorts to the tail of the preceding vowel for feature extraction, so that (time-varying) AR modeling be possible, [8]. The time - frequency analysis employed here, on the other hand, can expose in every detail the different evolution of the frequency contents of the three unvoiced stops with time. It is therefore expected to provide features appropriate for discriminating among the three phonemes.

## 2. THE CLASSIFICATION SCHEME

The method proposed here for classification of unvoiced stop consonants is outlined in the block diagram of fig-

0-7803-2431-5/95 $4.00 

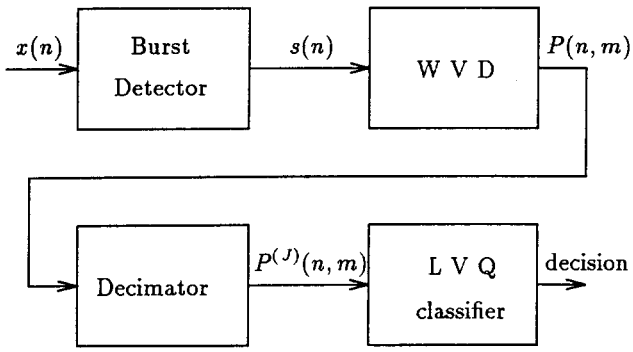


Figure 1: Block diagram for classification

ure 1, where the first three blocks comprise the feature extraction step, while the fourth one is the actual classification step.

### 2.1. Feature Extraction

As explained in the introduction, we have chosen to focus on the burst rather than the closure part of the stop phonemes for feature extraction. We have therefore employed a *burst detector* initial step, in order to extract from the initial TIMIT signals $x(n)$ a segment $s(n)$ containing the burst part of the stop phoneme. A simple energy detector was adequate at this point, thanks to the clear silence part preceding the burst. The duration of signal retained after burst was empirically set equal to 50 msec.

The time-frequency representation of $s(n)$ is computed in the next step. The Wigner-Ville distribution (WVD) has been selected, because all other Cohen class time - frequency distributions can be obtained from smoothed versions of it; it will therefore produce the "sharper picture" among them. For a discrete-time signal $s(n)$, this is given by the two dimensional function $W(n,\theta)$ as

$$W(n,\theta) = 2\sum_{k=-\infty}^{\infty} s(n+k)s^{\star}(n-k)\,exp(-j2k\theta), \quad (1)$$

where $n$ is the time variable and $\theta$ is the frequency variable, lying in [0, $\pi$], [1]. The *pseudo*-Wigner distribution (PWD) rather than the Wigner distribution itself is computable in practice from a finite-length signal, as

$$P(n,m) = 2\sum_{k=-L+1}^{L-1} w(k)s(n+k)\, w^{\star}(-k)s^{\star}(n-k)exp(-jk\frac{2\pi m}{M}), \quad (2)$$

where window $w(k)$ is $2L-1$ points long and $m = 0,1,\ldots,M-1$ is now the frequency point index, producing $M$ equispaced points in [0, $\pi$]. Either the resulting two-dimensional field by itself, or a set of its morphological characteristics, such as contours or dominant peak positions, can serve as the classification pattern. The former choice was made here.

The two-dimensional field $P(n,m)$ just computed is next smoothed using a convolutional window along the time axis only, and subsequently subsampled at an appropriate rate. This step has the twofold effect of (i) reducing the size of $P(n,m)$ and, consequently, of the feature vector used, and (ii) performing a time-averaging which is desirable, since the burst part of the signal is random while the computation of the WVD as in equation 1 does not provide any averaging by itself.

In fact, the smoothing - subsampling step can be performed at once using a multiresolution approach. Indeed, in multiresolution analysis, each approximation level produces a coarser version of its input, by low-pass filtering it and then by subsampling the output of the filter. This is what the decimator part produces. What we propose here is to pass the $P(n,m)$ two-dimensional field repeatedly through the decimator, along the time-axis only, i.e., successively for all frequency points $m = 0,1,\ldots M-1$, as follows:

$$P^{(j)}(n,m) = \sum_{k} h(2n-k)\; P^{(j-1)}(k,m), \quad (3)$$

where $j = 1,\ldots,J$ denotes the successive approximation levels. Initialization at $j=0$ is obtained by the raw output of equation 2, $P^{(0)}(n,m) = P(n,m)$. The filter $\{h(n)\}$ employed can be either the low-pass analysis filter of a perfect reconstruction filter bank (PRFB), or any low-pass filter with good cut-off behavior. Furthermore, this step removes a significant portion of the noise present in $P(n,m)$, through the repeated low-pass filterings it involves. The $J$-th level approximation, $P^{(J)}(n,m)$, is vectorized and fed into the classifier for training or testing.

### 2.2. Classification

The fourth step in the proposed method is the actual classification step. The LVQ classifier of Kohonen, [5], was selected because of its property to perform successful classification when the pattern space is not *linearly* separable into target classes. The latter is the case in our pattern space, as there is no evidence that the three target classes /k/, /p/ and /t/ are linearly separable, either in the pattern space of ours or of others, [8].

The LVQ classifier is initialized by a set of reference vectors in the pattern space, $\mathbf{r}_n, n = 1,2,\ldots,N$, either

randomly or through some other clustering algorithm. During the training phase, the set of training feature vectors, $\mathbf{t}_i, i = 1, 2, \ldots, T$, is repeatedly used to update the positions of $\mathbf{r}_n$'s, so that finally the boundaries they define are close to being optimal in the Bayesian sense.

During one such repetition or *epoch*, each training vector $\mathbf{t}_i$, $i = 1, 2, \ldots, T$, is assigned to the target class of one of the current $\mathbf{r}_n$'s by the 1-Nearest Neighbor (1-NN) rule. A misclassification (along with two other conditions) triggers a position update of the reference vectors involved (the correct one, $\mathbf{r}_c$, and the (wrong) nearest neighbor, $\mathbf{r}_w$), by the following rules: $\mathbf{r}_c$ is moved towards $\mathbf{t}_i$, while $\mathbf{r}_w$ is moved away from $\mathbf{t}_i$, as

$$\begin{aligned} \mathbf{r}_c(i) &= \mathbf{r}_c(i-1) + \alpha(i)\{\mathbf{t}_i - \mathbf{r}_c(i-1)\}, \\ \mathbf{r}_w(i) &= \mathbf{r}_w(i-1) - \alpha(i)\{\mathbf{t}_i - \mathbf{r}_w(i-1)\}, \end{aligned} \quad (4)$$

where step size $\alpha(i)$ must decrese monotonically with "time" $i$ to achieve convergence. The training procedure just described corresponds to what is known as the LVQ2 classifier, [6].

The reference vectors $\mathbf{r}_n, n = 1, 2, \ldots, N$, obtained during the final epoch of the training phase, represent the target classes. During the classification phase, they are used to classify an incoming test feature vector $\mathbf{v}$ by the 1-NN rule.

## 3. EXPERIMENTAL RESULTS

The unvoiced stop signals used to evaluate the proposed classification scheme were obtained at random from the TIMIT speech database, without any speaker/sex/context preference. They consist of 88 /k/'s, 42 /p/'s and 85 /t/'s, 215 in all. These were next passed through the burst detector, and the first 50 msec after burst, denoted by $s(n)$ in figure 1, were retained for further processing. At an 8 KHz sampling rate, this produced 400-points-long data records. The pseudo-Wigner-Ville distribution of each data record was then computed at $M = 16$ equispaced frequency points in $[0, \pi]$, but only half of them were retained thanks to symmetry. The resulting 2-D fields $P(n, m)$ were then replaced by their level $J = 7$ coarser approximations. The FIR filter $h$ = [0.0094, -0.0707, 0.0694, 0.49, 0.49, 0.0694, -0.0707, 0.0094], which is the low-pass analysis part of a PRFB quadruple, taken from [2], was used for this purpose. The resulting smoothed fields, denoted by $P^{(J)}(n, m)$ in figure 1, were next vectorized and trailing zeros were cut off, to yield finally 215 feature vectors of dimension $50 \times 1$ in the feature vector pool. In figures 2, 4 and 6 are shown three typical signals (closure and burst) of /k/, /p/ and /t/, respectively, while in figures 3, 5 and 7 are shown the corresponding 2-D classification patterns.

| | /k/ | /p/ | /t/ | Total |
|---|---|---|---|---|
| Test 1 | 90.0 % (53/59) | 87.5 % (28/32) | 100 % (59/59) | 94.0 % (140/150) |
| Test 2 | 93.1 % (27/29) | 50.0 % (5/10) | 100 % (26/26) | 90.0 % (58/65) |

Table 1: Classification scores

For the classification step, this feature vector pool was divided into (1) the training set, comprised of 59 /k/'s, 32 /p/'s and 59 /t/'s, 150 in all, chosen at random from the pool, and (2) the test set, comprised of all feature vectors not in the training set, which is 29 /k/'s, 10 /p/'s and 26 /t/'s, 65 in all. The LVQ2 classifier was trained using $\alpha(0) = 0.1$, 2000 epochs, and 35 neurons in the competitive layer, as training parameters. Correct classification scores on the training set, used as Test set #1, and on the actual test set, used as Test set #2, are shown in table 1. It can be seen that the results are satisfactory, as they compare favorably to analogous results of previous works. The slight only degradation in performance from Test #1 to Test #2 shows the good generalization property of this network. The poorer classification scores belong to class /p/, which is under-represented in the feature pool, and in the training set, consequently. This point suggests that balanced instead of random selection of the class representatives for training set would give higher classification scores. Higher scores are also expected if the training set is further divided into more than three target subclasses, exploiting contextual information. Other points that can be further tunned for higher classification scores are the training parameters, the version of the LVQ classifier employed, and the algorithm chosen to initialize it. Yet, the results obtained here on a limited size experiment show the potential of the proposed approach.

## 4. CONCLUSIONS

A classification scheme for unvoiced plosive consonants is proposed in this work. It performs time-frequency analysis of the burst part present in these phonemes to obtain classification patterns that suit their non-stationary and aperiodic nature. Classification with the LVQ classifier is performed on speech segments obtained from the TIMIT speech database with encouraging results under context- and speaker- independent conditions. Further research is carried out on the selection of the most suitable feature and on the type of classifier employed.

## 5. REFERENCES


[1] T.A.C.M.Claasen, W.F.G. Mecklenbrauker, "The Wigner Distribution - A tool for Time-Frequency analysis," *Philips J. Res.*, vol. 35, Parts I,II and III, 1980.

[2] R.E.Crochiere, L.R.Rabiner, "Multirate Digital Signal Processing," *Prentice-Hall,* New Jersey, 1983.

[3] A.K.Datta, N.R.Gangulli, S.Ray, "Recognition of unaspirated plosives - A statistical approach," *IEEE Trans. ASSP,* vol. 28, No 1, Feb. 1980.

[4] R.O.Duda, P.E.Hart, "Pattern Classification and Scene Analysis," *John Wiley & Sons,* New York, 1973.

[5] T.Kohonen, "Self-Organization and Associative Memory," 2nd Ed., *Springer-Verlag,* New York, 1988.

[6] T.Kohonen, "Improved Versions of LVQ," *Proc. IJCNN'90,* vol. 1, pp. 545-550, June 1990.

[7] K.F.Lee, H.W.Hon, R.Reddy, "An Overview of the SPHINX speech recognition system," *IEEE Trans. ASSP,* vol. 38, no. 1, Jan. 1990.

[8] K.S.Nathan, H.F.Silverman, "Time-Varying feature selection and classification of unvoiced stop consonants," *IEEE Trans. on Speech and Audio Proc.,* vol. 2, No. 3, pp.395-405, July 1994.

[9] L.R.Rabiner, R.W.Schafer, "Digital processing of speech signals," *Prentice-Hall,* 1978.

[10] M.Rangoussi, G.Carayannis, "Adaptive detection of noisy speech using third-order statistics," *Intl. J. Adapt. Control & Signal Proc.,* Wiley, 1994(to appear).


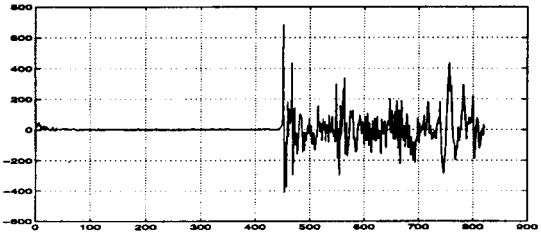

Figure 2: Speech signal, /k/.

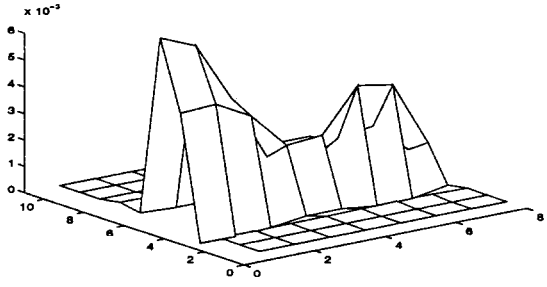

Figure 3: Smoothed WVD of /k/.

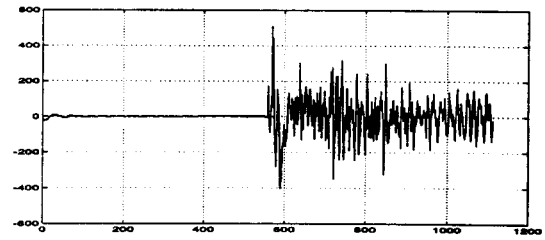

Figure 4: Speech signal, /p/.

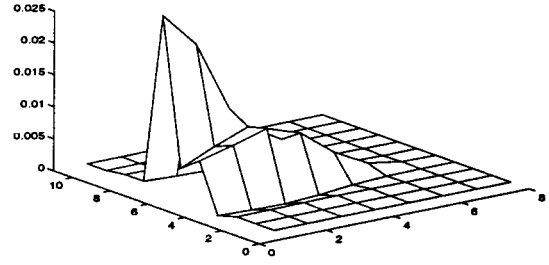

Figure 5: Smoothed WVD of /p/.

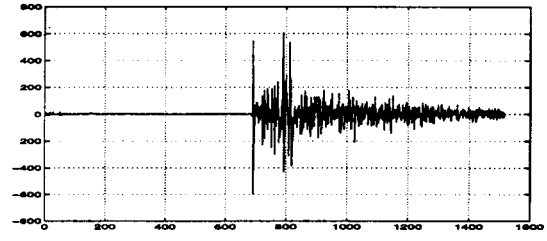

Figure 6: Speech signal, /t/.

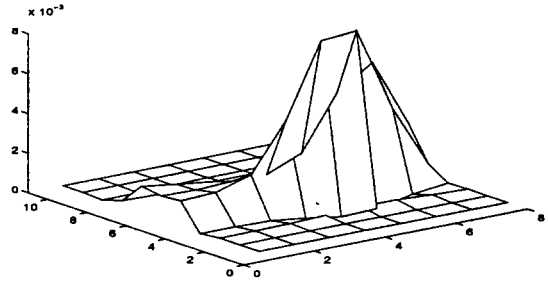

Figure 7: Smoothed WVD of /t/.