Plosive Spotting with Margin Classifiers

Joseph Keshet[1], *Dan Chazan*[2] *and Ben–Zion Bobrovsky*[1]

[1]Department of Electrical Engineering,
Tel Aviv University, Tel Aviv 69978, Israel
{keshet,bobrov}@eng.tau.ac.il

[2]HRL Audio/Video Technologies Group,
IBM Israel - Science and Technology, Haifa, 31905, Israel
chazan@il.ibm.com

Abstract

This paper presents a novel algorithm for precise spotting of plosives. The algorithm is based on a pattern matching technique implemented with margin classifiers, such as support vector machines (SVM). A special hierarchical treatment to overcome the problem of fricative and false silence detection is presented. It uses the loss-based multi-class decisions. Furthermore, a method for smoothing the overall decisions by sequential linear programming is described. The proposed algorithm was tested on the TIMIT corpus, which produced a very high spotting accuracy. The algorithm presented here is applied to plosives detection, but can easily be adapted to any class of phonemes.

1. Introduction

The plosives consonants (/b/, /d/, /g/, /k/, /p/ and /t/) are unique among phoneme categories in English since they involve three distinct stages which are sequential in time [1]:

1. **Closure (occlusion)** — The articulators totally block the air-stream and the air pressure increases just behind the obstruction. For the voiced plosives (/b/, /d/ and /g/), there is an underlying voicing activity during part of this stage.
2. **Burst** — The articulators quickly move away from each other. An explosive burst of air rushes through the opening, involving energy in most or all of the audible spectrum.
3. **Transition** — Transition segment to the next sound.

Nowadays, The HMM is the predominant acoustic model in continuous speech recognition systems. Inherently, the HMM suffers from three basic restrictions [2]: assumption of conditional independence of observations given the state sequence, features extraction imposed by framed-based observation, and duration model implicitly given by a geometric distribution. These restrictions result in a very poor model for plosives, and hence reduce recognition rates on this important class.

Several schemes have been proposed for special purpose plosives recognition machines, which were not based on HMM. Torres and Iparraguirre [3] proposed two knowledge-based classifiers for identification of Spanish unvoiced stops, which were designed and tested over a consonant-vowel (CV) context and resulted in a satisfactory rate of identification. Morris et al. [4] compared the baseline performance of human perception of the consonantal place of articulation with the performance of two automatic speech recognition techniques (Kohonen self organizing map and Gaussian mixture classifier) on multilingual VC and CV vocalic transition segments. Ali et al. [5] suggested a new set of acoustic features and a knowledge-based acoustic-phonetic system for automatic recognition of isolated stops, taken from continuous speech. Lin, Lee and Lin [6] presented methods for CV alignment of Chinese Mandarin speech, using fuzzy implication to find the abrupt spectral difference changes and spectral distance measuring. They reported their system performance was comparable to that of a human expert, though this system might fail to handle continuous English speech.

Although there have been some studies on segmenting out, and recognizing plosive within known patterns of speech (such as CV, VC, etc.), so far no work has been carried out on accurate segmentation and recognition of plosives in fluent speech. We propose a two stage scheme to carry out the recognition of plosives. During the first stage the exact location of the plosive is spotted, while in the second stage, the plosive is classified as a specific type, given its location [7]. It may be noted here that the purpose of the work is to obtain a model for plosives. While for clean speech the proposed two stage approach may be an effective scheme for actually recognizing the sound, for noisy speech a more global approach incorporating all the information about the speech data may be used, based on this model.

This paper focuses on the problem of precisely spotting the plosives within fluent speech. Particularly, we are interested in spotting the plosive burst, since it may serve as an anchor point to the identification process. Though the plosive transition to the succeeding phoneme may also be considered a reference point, it is often hard to spot accurately since the transition may be blurred. Note that both the burst and the transition are crucial to identification of the plosives for both humans [8] and machines.

2. Plosives Spotting with SVM

The plosive burst duration can vary from 10 msec to 50 msec, and its character may not be preserved under time scale modifications. Thus, when modelling plosives statistically, we must take into account their dynamic characteristics. One way of doing so is by constructing a statistical model for each possible burst length. Another way is by using a single model corresponding to the longest burst duration, i.e., 50 msec. The shorter bursts fit this this model with some possibly irrelevant data which often is of high energy and may mask the presenece of the plosive. For this reason it is important that the additional data be handled statistically by a properly constructed classifier which will make use of the relevant information which is present in the next sound and ignore the rest.

Let $\mathbf{x} = (x_1 \ldots x_n)^T$ denote an observation vector consisting of concatenations of acoustic features of several adjacent frames of speech. The convention adopted in this work was to have the first frame of the plosive coincide with the end of the plosive closure, while the remaining frames indicate the plosive burst, the transition and possibly part of the next phoneme. The number of frames within the observation vector is chosen to be fixed in such a way that it covers almost all possible durations of the plosive bursts. Let us assume that the observed vectors are drawn independently from a fixed but unknown probability distribution function, $p(\mathbf{x})$. These vectors lie in an n–dimensional metric space, $\mathcal{X}$. The metric which is used depends on the type of acoustic features. A label, $y \in \{-1, +1\}$, is attached to every vector in $\mathcal{X}$, according to the conditional distribution function $p(y \mid \mathbf{x})$, where $+1$ is assigned to vectors representing plosives, and -1 for all other vectors.

The problem addressed by this work is to find a function (classifier) $f(\mathbf{x}) \in \mathcal{F}$ in such a way that the sign of $f(\mathbf{x})$ estimates y. The choice of the correct function should be based on a training set of m independent, identically distributed observations drawn according to $p(\mathbf{x}, y) = p(y \mid \mathbf{x})p(\mathbf{x})$:

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m). \tag{1}$$

The training set contains examples of both plosives and all other phonemes, including transitions between phonemes which are not plosives.

One approach to classifying with some irrelevant features is the *empirical risk minimization* (ERM) principle: choose a function $f(\mathbf{x})$ from $\mathcal{F}$, which minimizing the *loss*, $L(y, f(\mathbf{x}))$, between the label y of a given vector $\mathbf{x}$ and $f(\mathbf{x})$, as is often done in the classic learning methods. Another approach is the *structural risk minimization* (SRM) principle, which prefers functions that minimize both the loss and the *VC dimension* (cf. [9]). The *support vector machine* (SVM) is a classifier designed to achieve the SRM goal. In this approach the generalization ability (smoothness) of the classifier is balanced out against its ability to fit the specific data.

To construct a SVM we map the original observations into some Hilbert space using a non-linear mapping, and then find an optimal decision hyperplane in that space. The non linear mapping may be chosen in a way which allows the decision function $f(\mathbf{x})$ to be described in the original feature space $\mathcal{X}$ by means of the hyperplane defined by α in the following manner [9].

$$f(\mathbf{x}) = \sum_{i=1}^{m} y_i \alpha_i \cdot k(\mathbf{x}, \mathbf{x}_i) + b, \tag{2}$$

where $\mathbf{x}$ is a new vector to be examined, k is a positive definite symmetric function, which called *the kernel*, and $\mathbf{x}_i$ are the elements from the training set. α_i and b can be found using the following quadratic programming problem:

$$\max_{\alpha} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \cdot k(\mathbf{x}_i, \mathbf{x}_j) \tag{3}$$

$$\begin{aligned} \text{subject to} \quad & 0 \le \alpha_i \le C, \quad i = 1, \ldots, l \\ \text{and} \quad & \sum_{i=1}^{m} \alpha_i y_i = 0. \end{aligned} \tag{4}$$

C is the trade-off between training error and $yf(\mathbf{x})$, which called *the margin*.

The Gaussian kernel turns out to be a convinient choice.

$$k(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|/\sigma^2},$$

This adds two more parameters to be adapted namely σ and C. These parameters are set iteratively, using the method suggested recently by Chapelle and Vapnik [10].

The results of the plosive spotting with the above method are presented in section 5 with detailed comments. Attention should be drawn to the fact that detection of fricatives and silences as plosives was common (see Table 2). In addition, multiple detections of the same plosive occured frequently. The former is handled by the *hierarchical* decision method described in the next section, and the latter is handled by *sequential decision smoothing* described in section 4.

3. Hierarchical spotting of plosives

Fant [1] had observed that after the plosive's burst a short fricative sometimes appears. This may explain the detections of fricatives as plosives. The false detections of silences is probably due to their similarity to the plosive's closure.

We add two classifiers: one which discriminates between fricatives and plosives and another one which discriminates between silences and plosives. One way to combine the three classifiers is to declare a plosive to have occured only if all classifiers (plosives versus all, plosives versus fricatives and plosives versus silences) agree that a plosive had occured. We call it the "logical-and" decision.

Let us describe this new setting as a pseudo multi-class problem with four classes (plosives, fricatives, silences and "all other phonemes") but with a binary decision (plosives or not). A general point of view for the multi-class problem was suggested by Allwein, Schapire and Singer [11] and Crammer and Singer [12]. The idea is to associate each class r with a row of a *coding matrix* $\mathbf{M} \in \{-1, 0, +1\}^{k \times l}$, where k is the number of classes and l is the number of classifiers involved. Each column of $\mathbf{M}$ describes a binary hypothesis f_i, for which +1 and -1 indicate the labels of the classes involved, and 0 indicates we don't care how hypothesis f_i categorizes examples from this class.

The setting of the coding matrix $\mathbf{M}$ of our problem is as follows:

$$\mathbf{M} = \begin{pmatrix} +1 & +1 & +1 \\ -1 & -1 & 0 \\ -1 & 0 & -1 \\ -1 & 0 & 0 \end{pmatrix},$$

where the first column corresponds to the classifier of plosives versus all other phonemes (including fricatives and silences), the second column corresponds to the classifier of plosives versus fricatives (we don't care about silences and other phonemes) and the third column corresponds to the classifier of plosives versus silences (we don't care about fricatives and all other phonemes).

The decision rule, based on the coding matrix, is computed as follows. Let $\mathbf{M}(r)$ denote the row r of $\mathbf{M}$, and let $\mathbf{f}(\mathbf{x}) \in \mathcal{F}^l$ be the vector of the classifier results of a test instance $\mathbf{x}$. We associate the label r to $\mathbf{x}$ in the following way [11]:

$$\arg\min_{r} d(\mathbf{M}(r), \mathbf{f}(\mathbf{x})), \tag{5}$$

for some distance d.

One way of choosing d is as *Hamming distance*, that is counting up the number of positions in which the sign of the vector $\mathbf{f}(\mathbf{x})$ differs from the row $\mathbf{M}(r)$, respectively.

Another approach of choosing d, the *loss-based approach*, is to set d as the total loss of the proposed class r. This approach is preferred, because it exploits the classifier's level of confidence in the prediction:

$$d(\mathbf{M}(r), \mathbf{f}(\mathbf{x})) = \sum_{i=1}^{m} L(\mathbf{M}(r,i), f_i(\mathbf{x})). \qquad (6)$$

Note that the loss for SVM is [13]:

$$L(y, f(\mathbf{x})) = \max\{1 - y \cdot f(\mathbf{x}), 0\}, \qquad (7)$$

therefore:

$$\arg\min_{r} \sum_{i=1}^{m} \max\{1 - \mathbf{M}(r,i) \cdot f_i(\mathbf{x}), 0\}. \qquad (8)$$

Empirical comparison between all the hierarchical methods presented above can be found in section 5.

4. Sequential Decision Smoothing

In order to overcome multiple detections of the same plosive, we post-process the decisions vector. The idea is to work with a window of length n_w, which is much larger than the length of the observation vector, n. The smoothing is based on some optimization with constraints over this window. Some possible constraints could be derived, for example, from bounds on the minimal time period between two consecutive plosives, the maximal duration of plosives or the permitted number of detections per window.

We used the following constraint: "leave only the first detection within the window and ignore all other detections". n_w was taken to be $2n$, that is, the window included at most one plosive.

Let $\{\hat{y}_i\}_{i=1}^{n_w}$ be the individual decisions making up the decisions vector. We state the optimization problem as follows:

$$\max_{\{\hat{y}_i\}} \quad \sum_{i=1}^{n_w} (n_w - i)(1 + \hat{y}_i) \qquad (9)$$

$$\begin{aligned} \text{subject to} \quad & \sum_{i=1}^{n_w} (1 + \hat{y}_i) \le 2, \\ \text{and} \quad & \hat{y}_i \in \{-1, +1\}. \end{aligned} \qquad (10)$$

This optimization problem can be solved using integer programming techniques.

5. Results

We used the TIMIT corpus to build the training sets for the three classifiers, since it has time-aligned phonetic transcriptions. The training sets included 12200 plosives, 10000 fricatives, 2800 silences and 65000 other phonemes. Each entry in each of the training set was a vector of acoustic features concatenation of several frames. The acoustic features were the first coefficients of the MFCC plus the log energy, based on ETSI standard for distributed speech recognition front-end [14].

Table 1 summaries the set of features for each classifier. The last row indicates the number of MFCC coefficients per single frame (including the log energy). We selected high time-resolution frame lengths (5 msec) and low frequency-resolution (first 5 MFCC) for discrimination of plosives against all other phonemes, since it should capture the sudden temporal changes of the plosive burst. Though, when discriminating plosives from fricatives we preferred lower time-resolution (10 msec) and higher frequency-resolution (first 10 MFCC) to avoid recognition of abrupt changes within the fricative as the plosive burst. Slightly more careful treatment is needed for distinguishing plosives from silence features, where we added 50% frame shifting.

Table 1: *Set of features of each classifier.*

	plosives vs. all	plosives vs. fricatives	plosives vs. silences
frames per vector	8	4	8
frame length [msec]	5	10	5
frame shift [msec]	5	5	2.5
MFCC per frame	5	10	5

We define *insertions* as the percentage of the number of false frame detections relative to the number of all non-plosives frames. We also define *deletions* as the number of miss-detected plosives relative to the total number of plosives. Note that the decisions were taken per frame, and we allowed deviation of two frames (10 msec) from the TIMIT segmentation for the decision to be counted as correct.

We compared the detection error rate for all methods presented above using a DET curve [15]. The detection line was created by varying a threshold on the function $f(\mathbf{x})$ of each of the margin classifiers involved.

The results for the TIMIT core test set are presented in Figure 1, for each of the hierarchical method described above, and after sequential smoothing. For comparison, we placed on the figure the detection error rate of a commercial large vocabulary continuous parameter HMM system with 3000 context dependent states and 10 components per state trained on 500 hours of speech (not TIMIT) operating as phonetic classifier with the same method of error computation described above.

We present in Table 2 the distribution of insertions among phoneme classes. The analysis corresponds to a point on the DET curve with deletion of 1.2% and insertion 2%, with neither hierarchical improvements nor with sequential smoothing.

Table 2: *Insertions distribution among phoneme classes (at the detection point with deletion of 1.2% and insertion 2%).*

Phoneme class	Insertions [%]
silences	34
fricatives	28
vowels	23
glides	7
nasals	6.5
affricates	1.5

It appears that the hierarchical method outperform the plain classification setup, especially for insertions. This seems reasonable since we added two classifiers which handle specifically the most problematic insertions.

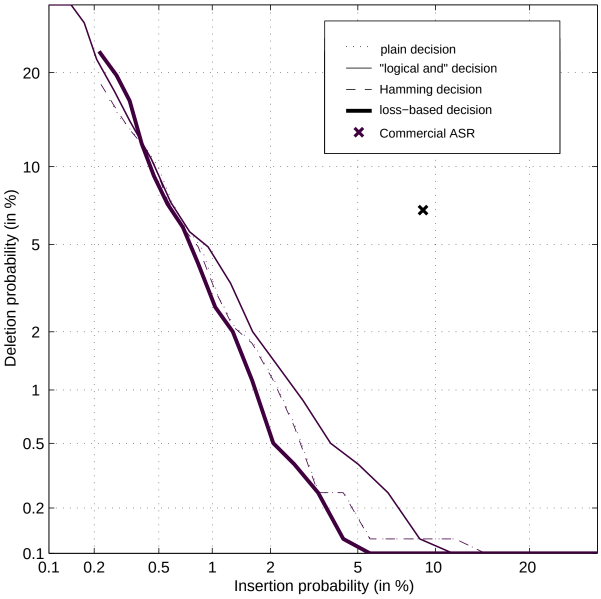

Figure 1: *Detection error tradeoff curves of plosive spotting for different decision methods.*

6. Discussion

We presented an algorithm for accurate spotting of the plosives which was based on large margin classifiers, such as the SVM. We showed a method to reduce the false-alarms (insertions) by treating the multi-class problem, and making hierarchical decisions. Finally, we introduced the smoothing of the decisions by integer programming.

Further research should be carried out on integration of the plosive segmentation system presented here and one of the plosive classification systems presented in the introduction into one system. Furthermore, the system presented here is for plosives, but can easily be adapted to any class of phonemes.

The experimental results show that the accuracy of spotting the plosives using these methods is very high, compared to an HMM based recognizer, and indicates a great potential to improve the HMM based systems. This can be done, for example, by forcing the the Viterbi search to pass at plosives states when found.

7. Acknowledgement

Thanks are due to Yoram Singer for pointing out the references on optimal binary decision based on sub class discriminators and the smoothing procedure used to select the actual plosive location.

8. References

[1] G. Fant, “Stops in CV syllables,” in *Speech Sounds and Features*, pp. 111–139, MIT Press, Cambridge, MA, 1973.

[2] M. Ostendorf, “From HMMs to Segment Models: Stochastic Modeling for CSR,” in *Automatic Speech Recognition - Advanced Topics*, pp. 185–210, Kluwer Academic Publishers, 1996. C. H. Lee, F. K. Soong and K. K. Paliwal, eds.

[3] M. I. Torres and P. Iparraguirre, “Acoustic parameters for place of articulation identification and classification of spanish unvoiced stops,” *Speech Comm.*, vol. 18, pp. 369–379, 1996.

[4] A. C. Morris, G. Bloothooft, W. J. Barry, B. Andreeva and J. Koreman, “Human and machine identification of consonantal place of articulation from vocalic transition segments,” in *Proc. Eurospeech*, pp. 2123–2126, 1997.

[5] A. M. Ali, J. V. Spiegel and P. Mueller, “Acoustic phonetic features for automatic recognition of stop consonants,” *J. Acous. Soc. Amer.*, vol. 103, no. 5, pp. 2777–2778, 1998.

[6] M.-T. Lin, C.-K. Lee and C.-Y. Lin, “Consonant/vowel segmentation for Mandarin syllable recognition,” *Comp. Speech and Lang.*, vol. 13, pp. 207–222, 1999.

[7] Z. Litichever and D. Chazan, “Classification of transition sounds with application to automatic speech recognition,” *submitted to Proc. Eurospeech*, 2001.

[8] B. H. Repp and H.-B. Lin, “Acoustic properties and perception of stop consonant release transients,” *J. Acous. Soc. Amer.*, vol. 85, no. 1, pp. 379–396, 1989.

[9] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1999.

[10] O. Chapelle and V. N. Vapnik, “Choosing kernel parameters for support vector machines,” *submitted to Machine Learning*, 2000.

[11] E. L. Allwein, R. E. Schapire and Y. Singer, “Reducing multiclass to binary: A unifying approach for margin classifiers,” *J. of Machine Learning Research*, vol. 1, pp. 113–141, 2000.

[12] K. Crammer and Y. Singer, “On the learnability and design of output codes for multiclass problems,” in *Proc. of the Thirteen Annual Conference on Computational Learning Theory (COLT)*, pp. 35–46, 2000.

[13] B. Schölkopf, A. J. Smola, R. Williamson and P. Bartlett, “New support vector algorithms,” Technical Report NC2-TR-1998-031, NeuroCOLT2, 1998.

[14] “Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms,” ETSI Standard ETSI ES 201 108, ETSI, Apr. 2000.

[15] A. Matrin et al., “The DET curve in assessment of detection task performance,” in *Proc. Eurospeech*, pp. 1895–1898, 1997.