Voiced-Unvoiced-Silence Classification of Sounds in a Speech Signal

M.G. Di Benedetto
Universita' degli Studi di Roma

Abstract

In this paper, a bayesian approach is described for deciding whether a given segment of speech signal should be classified as voiced speech, unvoiced speech, or silence. For this purpose, four measurements are made on each segment and, on the basis of the former, the a posteriori probabilities of the three classes are determined. The a priori probabilities of the classes are up-dated at each stage, by assuming that the sequence of the classes constitutes a Markov series, which is stationary and of the first order. This method has been found to provide reliable classification with speech segments as short as 12.8 ms and has been used in phoneme recognition applications.

Introduction

The problem of determining voiced speech (V), unvoiced speech (UV), and silence (S) arises in many speech analysis systems. In the past, a number of methods have been suggested for the V/UV/S classification. A few are of a statistical nature, while others are deterministic. For instance, Atal and Rabiner [1] make use of the maximum likelihood criterion, based on a statistical model. Rabiner and Sambur [2] propose that two distances be combined in a non linear way, in order to obtain the classification function. On the other hand, Siegel [3] suggests a deterministic method : by means of a linear programming algorithm, a discrimination function is determined which allows the differentiation among the three classes V/UV/S.

Various methods for the V/UV/S decision work in conjunction with pitch analysis such as the cepstral pitch detector presented by Noll [4]. One must note, however, that, for some applications, such as speech segmentation or speech recognition, the relation between the V/UV decision and the pitch extraction unnecessarily complicates the issue, especially at the boundaries between voiced and unvoiced segments.

In this paper, a bayesian approach is presented for deciding whether a speech segment should be voiced, unvoiced, or silence. For this purpose, at each stage, some measurements are made and a decision is taken for that class the a posteriori probability of which is maximum. Under the hypothesis that the probability density function (pdf) of the measurements is gaussian, the decision will depend upon the a priori probability, the mean vector, and the covariance matrix of each class. The sequence of the classes is considered to be a Markov series which is stationary and of the first order. In this manner, it is possible, at step k, to update the a priori probabilities of the classes which intervene in the computation of the a posteriori probabilities of the classes at the next step (k+1). Under the condition that the sequence of the classes is a Markov series of the first order, it is admitted that the decision at step (k+1) depends somehow upon the decision at step k. This intuitive approach appears logical in accordance with the experimental data obtained.

The extraction of the features is obviously of fundamental importance. The following features have been chosen, for reasons which will be explained in the paragraph touching upon the experimental results.

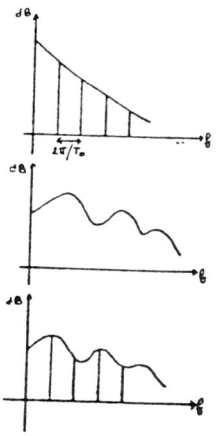
1. Energy of the signal
2. Zero-crossing of the signal
3. Autocorrelation coefficient at unit sample delay
4. First predictor coefficient

The paper is organized as follows: the second part deals with speech production; in the third section, the problem of how V/UV/S recognition is connected to the problem of phoneme recognition and the pattern in which this research has been developed will be described; finally, in the fourth part the algorithm of the V/UV/S classification and the experimental results obtained from its application will be indicated.

The sounds of speech and their production

Let us first define the characteristics of voiced sounds and those of unvoiced sounds, in order to clarify the meaning of these terms. In the first place, the apparatus for the production of speech will be considered as a linear system which can be characterized by its transfer function.

When the air crosses the glottis and provokes vibrations of the vocal chords, the glottal signal is pseudo-periodic and the sound emitted is called voiced.



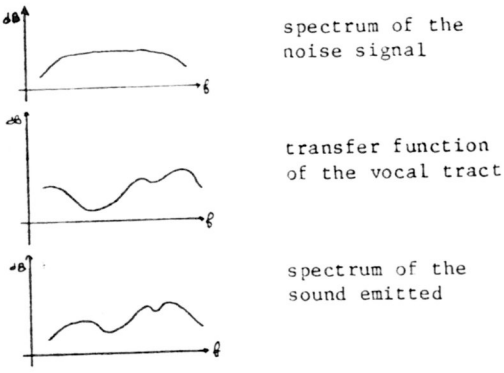
spectrum of the glottal signal

transfer function of the vocal tract

spectrum of the sound emitted

Fig.1 Voiced speech production

TT0115-6/82/0000-0400 $00.75



spectrum of the noise signal

transfer function of the vocal tract

spectrum of the sound emitted

Fig.2 Unvoiced speech production

this case, the fundamental frequency of the glottal gnal, which is variable according to various factors ch as the tension of vocal chords, is called the ndamental frequency of speech or pitch and is sponsible for the height of sound emitted. In this se, in which the vocal chords do not vibrate, the und emitted is called unvoiced (see Fig.1 and Fig.2).

The linguistic entities which compose the various unds which constitute a language are called phonemes.

Presentation of the problem and its treatement

The new algorithm for the classification of V/UV/S ggested in this paper represents a first step for taining an identification system of phonemes, in view the automatic assistance to lip reading.

The solution of the problem of identifying ntinuous speech offers various fields of application. the present chapter, this problem will be analyzed riefly and the difficulties related to its resolution ill be highlighted, in addition to illustrating the attern in which this work has been developed and, herefore, the final goal for its solution.

he problem of speech recognition

A system for the recognition of isolated words epresents a great simplification of the problem, ecause the number of words is limited and the oundaries of the words are known. In most cases, the ords are considered as unbisected patterns, obtained, or example, by spectral analysis. The identification f a pattern is then a problem of recognizing, choosing mong all the patterns of the dictionary, that which is loser to the one to be considered, according to a istance to be defined.

However, an identification system of isolated ords does not permit very rich communication with a omputer. In fact, one is obliged to pronounce the entences word by word.

In order to give a natural aspect to the dialogue, ne must pass to continuous speech. Difficulties arise ue to the following facts:

1. The definition of phonemes is physically not precise nd is not expressed in a precise mathematical mode as a function of measurements of the speech signal.
2. The characteristics of phonemes vary considerably rom speaker to speaker and also from an instant to nother for the same speaker.
3. In the continuous speech, there are no boundaries.
4. There is imprecision on the segmentation of speech in phonemes and also on the identification of the latter.

Help for lip reading

The project in which one has been involved studies, in collaboration with an institute for deaf children, the application of informatics techniques to the problems of the deaf.

The principle on which lip reading is based is to distinguish the sounds of the language according to the movements of the lips. On the other hand, different sounds present the same lip-image. For example, /p/, /b/, and /m/ or /f/ and /v/ cannot be distinguished. The consequence is that lip reading is ambiguous and requires a strong effort. It is natural to believe that complementary information would eliminate the ambiguity and would contribute at the same time to the acquisition of the spoken language on behalf of deaf children.

Some progress in this direction has been achieved when the "Cued Speech" was invented. This method, suggested by Dr.R.Orin Cornett [5], succeeds in eliminating such an ambiguity. In such a method, each complement is, in fact, the combination of a shape and a position of the hand. The positions of the hand correspond to six groups of vowels and the shapes to six groups of consonants. The set of information which can be obtained from the movement of the lips, the shapes and positions of the hands, permits the precise identification of the consonant-vowel syllaba.

One should not confuse the "Cued Speech" with the sign language, as the latter is of a fundamentally different nature, because the signs are associated to concepts expressed in the spoken language and not to some sounds of the latter.

The "Cued Speech" demonstrates that the spoken language can be correctly perceived visually and constitutes a proof of the usefulness of the lip complements.

The problem consists in reproducing such complements automatically. This project has, as its long-term objective, the realisation of a portable apparatus, which would permit to observe the lips of the speaker and, at the same time, the lip complements provided in real time in the form of light signals.

One must therefore elaborate identification algorithms of phonemes at a reliable level which permit the production of such complements.

For this purpose, it appears essential to establish a classification of sounds according to their characteristics of voiced and unvoiced.

V/UV/S/ classification: a bayesian approach

One of the first steps for obtaining an identification system of phonemes, with the aim of producing auxiliary lip complements which we mentioned previously, consists in classifying the speech segments in the three classes V/UV/S.

Let us consider, in fact, the classification of the phonemes according to their importance as regards the information rate in language (see Fig.3) [6].

T	tdn
I	eɛ̃i
P	pbm
r	kgR
S	sz
O	oøɔ̃
V	fv
Y	yu
A	aã
E	œ̃ɛ
H	ʃʒ
O	ɔœ

Fig.3- Classification of phonemes.

This table shows how the V/UV/S information could permit the elimination of the ambiguity of certain phonemes such as /p/-/b/ or /t/-/d/ which belong to particularly important groups.

Type of proposed algorithm

The algorithm suggested is of the pattern recognition type. Briefly, a pattern recognition machine attributes a class to a specific variable, on the basis of physical measurements made on the latter, as shown in Fig.4.

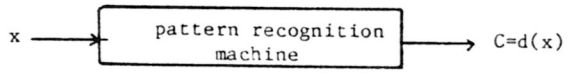


Fig.4- Pattern recognition

The pattern recognition machine can be considered as separated in two parts: a feature extractor and a classifier as shown in Fig.5.

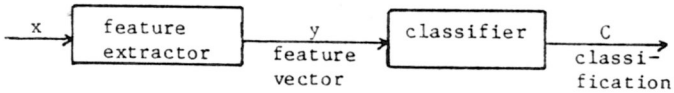


Fig.5- Feature extractor and classifier

The feature extractor reduces the dimension of the input vector to the classifier. The classifier discriminates x on the basis of y. As the dimension of y is less than x, the transformation produces a loss of information. The feature extractor should reduce the dimension, but, at the same time, it should maintain a high level of functionality. In the pattern recognition methods of a statistical nature, the classification element is a distance derived from a probability.

Model used to describe the phenomenon

As was mentioned previously, the algorithm presented is of the pattern recognition type. In particular, the classification element is bayesian. In order to clarify the questions which will be discussed later, it is essential to introduce the following notations:

x(k)	vector of measurements of the k^{th} segment
X(k)	aleatory variable for which the previous vector is a realisation
g(x;m,R)	pdf of x, multidimensional gaussian, for which the expected value is m, and the covariance matrix is R.
Pr(...)	probability of the event ...
C(j)	set of classes (j=0,1,2, S/V/UV/ respectively, to which the speech segments belong)
s(n)	class to which belongs the n^{th} speech segment
s*(n)	result of the classification of the n^{th} speech segment
p(i,j)	transition probabilities of the sequence of the classes
p(j)	a priori probabilities of the classes

Given a speech segment, let x=(f1,...,fN) be the vector of the measurements made. The vector x on which the k^{th} classification is based, before being observed, is an aleatory variable notated X(k). Suppose for simplicity that X(k) has a gaussian pdf of dimension N:

$$p_k(x/m,R) = g(x;m,R)$$

where m is the expected value and R is the covariance matrix (these quantities depend upon the class s(k)).

This hypothesis has been introduced in order to simplify the model. However, it is important to note that the method proposed does not depend upon the type of probability density function chosen.

Suppose, in addition, that the mean vector m and the covariance matrix R are deterministic variables. This hypothesis corresponds to the fact that these quantities are considered to assume the same values from speaker to speaker and therefore refer to the mean of the speakers. Let m(j) and R(j) be the values assumed corresponding to the class C(j), j=0,1,2. This is expressed by:

$$P_k(x/s(k)=C(j),m) = g(x;m(j),R(j))$$

Let the succession of the classes s(1), s(2),... , s(n),... be a Markov chain, which is stationary and of the first order. This hypothesis is motivated by the fact that a dependency exists betweeen two consecutive speech segments in the structure of the language. Naturally, it could be possible to increase the order of the Markov chain. Let p(i,j) represent the transition probabilities of the classes, i.e.:

$$p(i,j) = Pr(s(n+1)=C(j)/s(n)=C(j))$$

and p(i) the a priori probabilities of the classes C(i) (i,j=0,1,2,).

The quantities p(i,j), p(i), m(i), R(i), (i,j=0,1,2) must be estimated during the training phase.

Algorithm proposed for the classification at step n

Let

$$Y(n) = (x(1),x(2),...x(n)) = (Y(n-1),x(n))$$

be the vector which represents the series of measurements made on consecutive speech segments.
Let, in addition,

$$S(n) = (s(1),s(2),...s(n)) = (S(n-1),s(n))$$

be the vector which represents the sequence of the classes to which the speech segments, on which the measurements Y(n) have been made, belong. At step n, the class C(k) will be chosen so as to verify the condition:

$$Pr(s(n)=C(k)/Y(n-1),x(n)) = \max_j Pr(s(n)=C(j)/Y(n-1),x(n)) \qquad (1)$$

The class the a posteriori probability of which is maximum is then chosen as the true class.
One can also write

$$Pr(s(n)=C(j)/Y(n-1),x(n)) = \frac{Pr(s(n)=C(j)/Y(n-1)) \cdot p_n(x(n)/s(n)=C(j),Y(n-1))}{p(x(n)/Y(n-1))} \qquad (2)$$

where

$$p(x(n)/Y(n-1)) = \sum_{k=0}^{2} Pr(s(n)=C(k)/Y(n-1) \cdot p_n(x(n)/s(n)=C(k),Y(n-1)) \qquad (3)$$

$Pr(s(n)=C(j)/Y(n-1))$ indicates the a priori probability of the class C(j) when the measurements Y(n-1) are known. $p_n(x(n)/s(n)=C(j),Y(n-1))$ is the pdf of x(n) if the segment n belongs to the class C(j) and the Y(n-1) are known.

If one supposes that, when s(n) is known, X(n) is statistically independent of X(1),X(2),...,X(n-1), and takes into account that m is a deterministic variable, with value m(j), for the class C(j), one can write:

$$p_n(x/s(n)=C(j),Y(n-1)) = g(x;m(j),R(j))$$

Updating of the a priori probabilities of the classes

The classification at step (n+1) requires the updating of the quantity that appears at the numerator of Eq.2 : $Pr(s(n+1)=C(j)/Y(n))$ which represents the a priori probability of the class C(j).

As it is plausible to believe that the probabilistic structure of the language is independent of the physical realisation of sounds, one can suppose that:

$$Pr(s(n+1)=C(j)/S(n),Y(n))=Pr(s(n+1)=C(j)/S(n))$$

In addition, supposed that the succession of the classes S(n)=(s(1),...,s(n)) is a Markov series of the first order, one has:

$$Pr(s(n+1)=C(j)/S(n))=Pr(s(n+1)=C(j)/s(n))$$

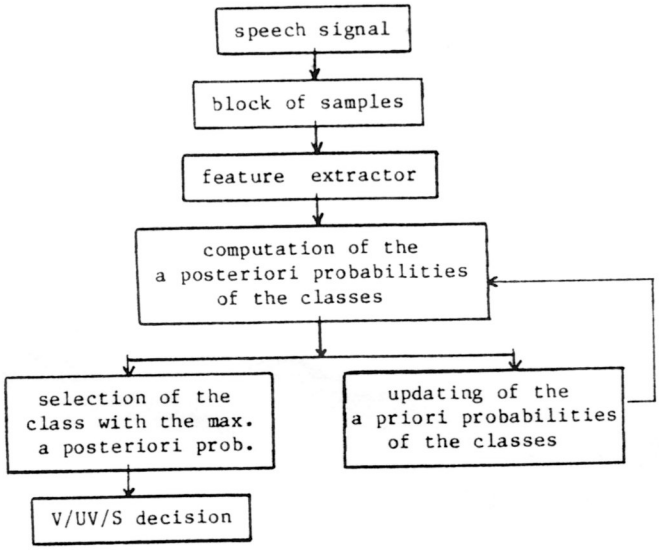


Fig 6- Algorithm proposed for the classification

Then the a priori probability of the class C(j) can be written as:

$$Pr(s(n+1)=C(j)/Y(n)) =$$

$$= \sum_{S(n)} Pr(s(n+1)=C(j),S(n)/Y(n))$$

$$= \sum_{S(n)} Pr(s(n+1)=C(j)/S(n),Y(n)).Pr(S(n)/Y(n))$$

$$= \sum_{i=0}^{2} p(i,j).Pr(s(n)=C(j)/Y(n)) \qquad (4)$$

Eq.4 gives the expression which must be computed for the updating of the a priori probabilities of the classes. The two terms that appeàr on the right of this equation are the transition probabilities between the classes p(i,j) and the a priori probabilities of the classes which have been calculated and used for the decision at the previous step. The algorithm is described in Fig.6.

Training

The classification criterion we have described requires the estimation of the initial values m(j), R(j) p(i,j) and p(j) i,j=0,1,2. This estimation is carried out on the basis of a set of sentences pronounced by different speakers.
Let the vector of the measurements made on the segment be:

$$x(n,j,i)=(x_1(n,j,i),...,x_L(n,j,i))$$

where

n=1,2,..,s	index of the speaker
j=0,1,2.	index of the class to which x belongs
i=1,2,..,M(n,j)	index of the vector of the classes C(j) of the n^{th} speaker
M(n,j)	number of segments of the speaker n classified C(j)

The transition probabilities p(i,j) are supposed to be independent of the speaker. Their estimation can then be obtained by calculating the frequencies:

$$p(i,j) = n(i,j)/d(j)$$

where n(i,j) is the number of times that the couple of the consecutive classes (C(i),C(j)) is observed and d(j) the number of times that the class C(j) is observed.

The probabilities of the classes are estimated by counting the number of times they are observed with respect to the total number of segments.

For the speaker n, the mean vector m(n,j) of the set of the vectors of class C(j) can be evaluated in the following manner:

$$m(n,j) = \sum_{i=1}^{M(n,j)} x(n,j,i)/M(n,j)$$

and the associated covariance matrix

$$r(n,j)= \sum_{i=1}^{M(n,j)} (x(n,j,i)-m(n,j))^t.(x(n,j,i)-m(n,j))/M(n,j)$$

An estimation for the value m(j) is given by

$$m(j) = \sum_{n=1}^{s} p(j,n).m(n,j)$$

where

$$p(j,n) = M(n,j)/MT(j)$$

$$MT(j) = \sum_{n} M(n,j)$$

Experimentation

Experimental conditions and features extraction

The algorithm described previously has been tested on an IBM Series/1 computer. The signal is filtered at 4.8 kHz by means of a low-pass filter (8 poles Butterworth) and then sampled at 10 kHz. A dynamic microphone has been used for the recording. The signal has been subdivided into blocks of 12.8 ms (i.e. 128 samples) and every segment obtained in this way is classified as V, UV or S.

Let N be the number of samples in a segment and let s(n) ,n=1,...,N, represent the speech signal. For every segment the following features have been extracted (s(0) indicates the last sample of the preceding window):

1. Log energy E defined as:

$$E = 10.\log[\sum_{n=1}^{N} s^2(n)]/N$$

2. Number of zero-crossings in the block which gives an idea of the localisation, in frequency, of the energy of the signal.
3. First linear predictor coefficient a(1). For the computation of a(1), a p=12 poles linear prediction analysis has been made, where the following expression has been minimized:

$$\sum_{n=1}^{N}[(s(n) + \sum_{k=1}^{p} a(k).s(n-k)]^2/N$$

a(k), n=1,..,N, are the linear predictor coefficients. The coefficient a(1) is in direct relation with the spectrum.
4. First normalized autocorrelation coefficient at unit sample delay:

$$C = \frac{\sum_{n=1}^{N} s(n).s(n-1)}{[\sum_{n=1}^{N} s^2(n).\sum_{n=0}^{N-1} s^2(n)]^{1/2}}$$

Several features have been extracted because a single feature is not sufficient to characterize the signal from the point of view V,UV,S. In fact, if one considers the energy, the latter appears strong if the signal is V, and weak if the signal is UV or S. In the same way, the first autocorrelation coefficient approaches one for a segment V or S, and zero for UV. For this reason, the classification is made combining the information of these various parameters.

In the training phase, sentences have been segmented and the segments obtained have been classified manually with standard procedures: examination of the spectrogram, waveform, etc. Doubtful segments were not taken into account.

In order to evaluate the performance of the algorithm, the results have been compaired to those obtained by using Atal and Rabiner's method [1], where, the vector of measurements is supposed to be gaussian, i.e.:

$$p_n(x/s(n)=C(j)) = g(x;m(j),R(j))$$

The V/UV/S decision is made on the basis of the pdf of x; the class i will be chosen such that:

$$g(x;m(i),R(i)) = \max_h g(x;m(h),R(h))$$

It is in fact an a priori bayesian decision.

Experimental results

In the experimentation of this algorithm the classes UV and S have been aggregated. The training for the estimation of the mean vector of x and the associated covariance matrix has been made on five sentences (1946 segments) pronounced by two female speakers (F1 and F2). In addition, the a priori probabilities and the transition probabilities of the classes have been evaluated on the basis of 4936 segments. The algorithm has been tested on three sentences: the first pronounced by one of the speakers of the training (F1), the second by a female speaker who was not included in the training (F3), the third by a male speaker (M1). The results obtained are the following (the numbers are in %, on the left side the results of the algorithm described, on the right those of Atal and Rabiner's algorithm).

1. Speaker F1

total number of segments=479
number of segments V =384
number of segments UV = 95

i\o	V	UV
V	99.7	0.3
UV	4.2	95.8

i\o	V	UV
V	96.9	3.1
UV	1.1	98.9

2. Speaker F3.

tot=480
V=367
UV=113

i\o	V	UV
V	96	4
UV	6.2	93.8

i\o	V	UV
V	98.4	1.6
UV	6.2	93.8

3. Speaker M1

tot=477
V=413
UV=64

i\o	V	UV
V	100	0
UV	7.8	92.2

i\o	V	UV
V	99.3	0.7
UV	2	98

The global results are the following:

tot=1436
V=1164
UV=272

i\o	V	UV
V	99.6	0.4
UV	5.9	94.1

i\o	V	UV
V	98.2	1.8
UV	3.7	96.3

Global error: method presented:1.4% A-R:2.1%

Conclusion

The algorithm for V/UV/S classification, which has been described in the previous chapter, is an analytic tool in a phoneme analysis system, the final goal of which is to present to the deaf a visual information that eliminates the ambiguity of lip image. The following remarks can be pointed out:

1. The results are very satisfactory. The error rate is less than the one which is obtained with a non-adaptative method such as the A-R method.
2. One must note that the classification is wrong in the transitions V/UV and UV/V. It is then useless to produce smoothing as required in other methods [1].

Since this work has been completed, some further research has been carried out touching upon the V/UV/S classification problem, i.e.:

1. A more sophisticated algorithm, in which not only the a priori probabilities of the classes are updated but also the pdf of x, has been studied and implemented. Obviously, this second algorithm has yielded better results than the present method.
2. The previous algorithm has then been experimented by considering the pdf of x as a Cauchy-like distribution. With such a hypothesis, the number of computations decreases considerably and the results obtained are still better than those refered to a non-adaptive method [1].
3. Another step has been tested: only three parameters have been used for the classification. The first autocorrelation coefficient has not been taken into account. As it was intuitive to believe, the error rate increases, but it is interesting to note that it is always less than the one obtained with a non-adaptative method [1].

It now appears interesting to extend the method for V/UV/S classification to the problem of phoneme recognition. For example, it should be possible to consider a number of classes corresponding to the number of phonemes, and try to characterize the latter by means of well chosen features.

References

[1] B.S.Atal/L.R.Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Application to Speech Recognition" -IEEE Trans. ASSP-24, No 3, June 1976.

[2] L.R.Rabiner/M.R.Sambur, "Application of an LPC Distance Measure to the Voiced-Unvoiced-Silence Detection Problem" -IEEE Trans. ASSP-25, No 4,1977.

[3] L.S.Siegel, "A Procedure for Using Pattern Recognition Techniques to obtain a Voiced-Unvoiced Classifier" -IEEE Trans. ASSP-27, No 1,1979.

[4] A.M.Noll, "Cepstrum Pitch Determination" -J.A.S.A. vol.41, pp 293-309,1967

[5] R.O.Cornett, "Cued Speech" -American Annals of the Deaf, vol.112, 3-13,1967

[6] M.D.Di Benedetto/F.Destombes/J.P.Tubach, "Utilisation de la theorie de l'information pour une etude quantitative de l'ambiguite en lecture labiale" -Proc. 12eme Journees sur la Parole, GALF, Montreal, 24,25,26 Mai 1981.

[7] J.Makhoul, "Linear Prediction of Speech" -Proc. IEEE, vol.63,pp 561-580, Apr.1975.

[8] J.P.Haton, "Reconnaissance de la Parole: l' Etat des Recherches" -01 Informatique No 331.

[9] T.Y.Young/T.W.Calvert, "Classification, Estimation and Pattern Recognition" -American Elsevier Publishing Company, Inc.