TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 15, No. 2, August 2015, pp. 368 ~ 372
DOI: 10.11591/telkomnika.v15i2.8374

Phonemes Classification Using the Spectrum

Ahmed El Ghazi*[1], Cherki Daoui[2]
Laboratory of Information Processing and Decision Support, Faculty of Sciences and Techniques, Béni Mellal, Morocco
*Corresponding author, e-mail: hmadgm@yahoo.fr[1], daouic@yahoo.com[2]

Abstract

In this work, we present an automatic speech classification system for the Tamazight phonemes. We based on the spectrum presentation of the speech signal to model these phonemes. We have used an oral database of Tamazight phonemes. To test the system's performances, we calculate the classification rate. The obtained results are satisfactory in comparison with the reference database and the quality of speech files.

Keywords: *phonemes, spectrum, gaussian mixture model tamazight*

1. Introduction

Some studies show that Automatic Speech Recognition (ASR) systems still lack performance when compared to human listeners in conditions that involve additive noise [1]. Such systems can improve performance in those conditions by additional levels of language and context modeling. However, this contextual information will be most effective when underlying phoneme sequence is sufficiently accurate. Hence, robust phoneme classification is a very important stage of ASR [2-4]. Accordingly, the front-end features must be selected carefully to ensure that the best phoneme sequence is predicted. In this paper, we investigate the performances of the speech spectrum and Gaussian mixture. Phoneme classification is commonly used for this purpose.

We are particularly interested in Moroccan Tamazight phonemes and we have selected the spectral features. For instance, the Mel Frequency Cepstral Coefficients (MFCC) is the most popular features used to model the speech signal. These features are the best modeling of the perception and production of human devices. In this work, we use the Gaussian model to classify the Tamazight phonemes. In this context, we took a population of phonemes that construct the digits from one to ten; we model each phoneme by features vectors. Phoneme classification by Gaussian mixture permits to collect the acoustic vectors that have the same characteristics. This classification can be used in hybrid with a hidden Markov model in particular applications [4-7]. The obtained classification rate is variable according to the phoneme type and its context.

Anyway, this paper is organized in the following manner. In section 2, we will give a description of speech spectrum; Section 3 describes a Gaussian mixture for a speech classification. Section 4 presents the experiments results. Finally, the study is ended by a conclusion.

2. Speech Spectrum

The speech spectrum is the presentation of signal on the three-dimensional space, the axis *X* presents a time, frequency in the axis *Y* and the level of each frequency in the axis *Z*. This analysis is obtained by using bunch of filters and Fourier transform. The following figure presents an example of a spectrum for the word 'yan' (number one). The black levels present a concentration of frequencies (Figure 1 and Figure 2).

Received April 17, 2015; Revised June 29, 2015; Accepted July 14, 2015

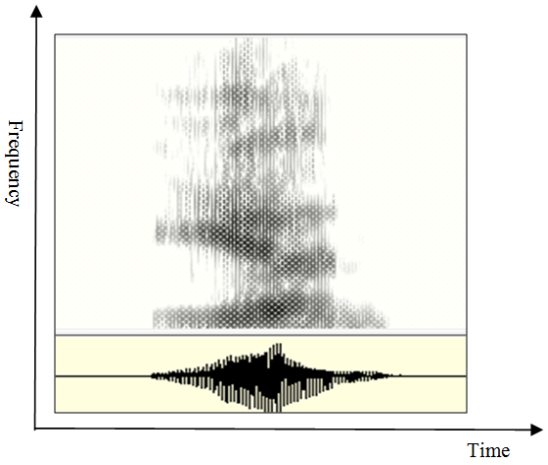

Figure 1. Speech spectrum for the word 'yan'

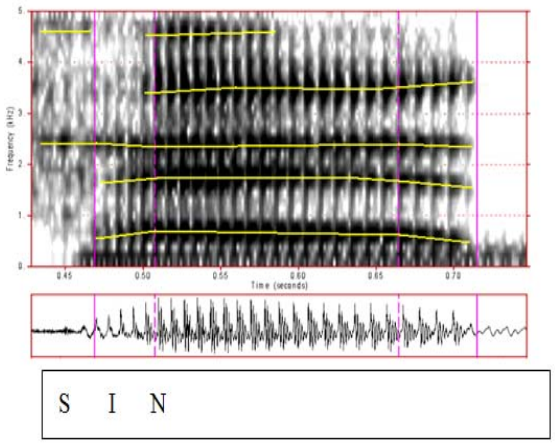

Figure 2. Spectrum of word 'SIN' and phonemes boundaries

This spectrum is variable according to the signal form. The speech spectrum is used to improve phonemes boundaries by detecting the areas of frequency concentration (Figure 2). This segmentation is approximate; it is difficult to determine the exact boundaries in speech signal.

3. Gaussians Mixture

The Gaussian mixture model (GMM) is an effective description of data sets comprising clusters of vectors that are more complex than simple Gaussian distribution. A Gaussian mixture model [1, 7], [9-11] is defined as:

$$\boldsymbol{f(x) = \sum_{i=}^{N} p_i g(x, \mu_i, \sum_i)}$$

Where $\boldsymbol{g(x, \mu_i, \sum_i)}$ is the Gaussian probability density function with mean $\boldsymbol{\mu_i}$ and covariance $\boldsymbol{\sum_i = \sigma_i^{jk}}$, *x* is a random D-dimensional vector, $\boldsymbol{x = (x^1, x^2, \dots, x^D)}$ and the $\boldsymbol{p_i}$ are weights which describe the relative likelihood of classes being generated from each of the clusters and must satisfy$\boldsymbol{\sum_{i=1}^{N} p_i = 1}$, where N is the number of classes.

In order to generate the GMMs from the phoneme training sequence, we employed the Expectation-Maximization (EM) algorithm [10], [12-13]. The EM algorithm for maximum-likelihood estimation of the parameters of a GMM is an iterative procedure in which each iteration consists of two steps: an estimation step (E-step), followed by a maximization step (M-step). In the E-step, the likelihoods, means and covariance matrix of GMMs are estimated depending on the observation sequence. In the M-step, the new values of the estimation of parameters of the GMMs are computed.

Suppose that we have a sample of S points $\boldsymbol{x_j = (x_j^1, x_j^2, \dots, x_j^D)}$, j=1,…, S, drawn from a set of points which are assumed to lie in N clusters. We initialize N Gaussians with probabilities $p_1=p_2=\dots =p_n=1/N$, means $\mu_1, \mu_2, \dots, \mu_n$, which can either be random or set equal to *N* of the data points with a small perturbation, and covariance matrices$\boldsymbol{\sum_1, \sum_2, \dots, \sum_n}$, set equal to the identity matrix [1].

In the E-step we compute:

The total likelihood:

$$\boldsymbol{t_j = \sum_{i=1}^{N} p_i g(x_j, \mu_i, \sum_i), j = 1, 2, \dots, S}$$

Where g is the Gaussian probability density function, the normalized likelihoods:

$$n_{ij} = p_i g(x_j, \mu_i, \sum_i) / t_j$$

The notional count:

$$C_i = \sum_{j=1}^{S} n_{ij}, i = 1, 2, \ldots, N$$

The notional means:

$$\bar{x}_i = \sum_{j=1}^{S} x_j n_{ij} / C_i, i = 1, 2, \ldots, N$$

And the notional sums of squares:

$$SS_i^{pq} = \sum_{j=1}^{S} x_j^p x_j^q n_{ij} / C_i, i = 1, 2, \ldots, N$$

In the M-step, we compute new values of parameters of the Gaussian model as follows:

$$p_i = C_i / S$$

$$\mu_i = \bar{x}_i$$

$$\sum_i^{pq} = SS_i^{pq} - \bar{x}_i^p \bar{x}_i^q$$

Where i=1, 2,…, N.

The Gaussian model can be assimilated to a hidden Markov model with one state (Figure 3). Each state represents a phoneme with *n* Gaussian components. The observations of this state are devided between the Gaussian components. In the classification step, we calculate the likelihood between the input features and all the Gaussian components.

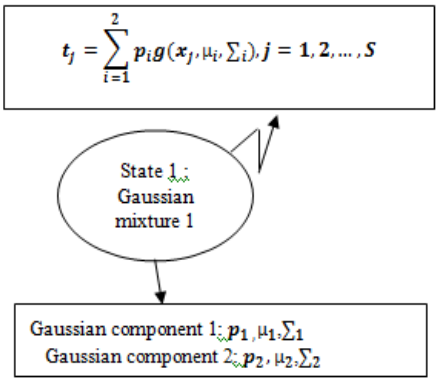

Figure 3. Gaussian mixture with 2 components

4. Database

The training data comprised a small vocabulary of ten isolated digits in Tamazight (from one to ten) spoken by ten speakers (5 males and 5 females) and test data spoken by five speakers. The produced signals are sampled at 16 KHz. Then, thespeech data was windowed (25 ms) and 512 pointsof FFTs were computed with a 256 points (12,5ms) advance between frames. The FFT coefficients were binned into12 Mel-spaced values to produce 12-dimentional featurevectors [1]. The Table 1 presents a training database and Table 2 presents the list of phonemes used.

Table 1. Training database

Numbers	Phonetic transcription	Tifinagh Transcription
1	Y A N	ⵢⴰⵏ
2	S I N	ⵙⵉⵏ
3	C R A DD	ⵛⵔⴰⴷⴷ
4	K O Z	ⴽⵓⵣ
5	S MM U S	ⵙⵎⵎⵓⵙ
6	SS DD E SS	ⵚⴷⴷⴻⵚ
7	SS A	ⵚⴰ
8	TT A M	ⵜⴰⵎ
9	T Z A	ⵜⵣⴰ

Table 2. Phonemes list used

Phoneme	Context	Symbol
/Y/	Y -A -N	ⵢ
/I/,/ N/	S -I-N	ⵙ, ⵉ
/C/,/R/	C -R -A -DD	ⵛ, ⵕ
/K/, /O/	K -O-Z	ⴽ, ⵓ
/S/, /U/	S-MM-U-S	ⵙ, ⵓ
/DD/	SS-DD-E-SS	ⴻ
/A/, /SS/	SS-A	ⵙ, ⴰ
/TT/, /M/	TT-A-M	ⵜ, ⵎ
/T/, /Z/	T-Z -A	ⵜ, ⵣ

5. Results of Phonemes Classification

The classification system permits to affect each phoneme to its Gaussian component. In the classification step, we calculate the likelihood between the phoneme featurein the input vectors and all the references Gaussians components. The obtained results are shown in Table 3.

Table 3. Obtained results

Phonemes	Classification rate	Error rate
/Y/	74%	26%
/I/	72%	28%
/C/	78%	22%
/K/	62%	38%
/S/	79%	21%
/DD/	80%	20%
/A/	75,5%	24,5%
/TT/	72,66%	27,34%
/T/	70%	30%
/N/	81%	19%
/R/	82%	18%
/O/	78,5%	21,5%
/U/	68%	32%
/SS/	70,5%	29,5%
/M/	72%	28%
/Z/	76%	24%

The error rate represents the classification error that illustrates the ambiguity between phonemes. At the acoustic level, there are common characteristics between speech units. The Table 4 presents some of these ambiguities.

Table 4. Some ambiguities between phonemes

phonemes	**i=/S/ j=/SS**	**i=/T/ j=/TT/**	**i=/O/ j=/U/**	**i=/C/ j=/K/**
Ambiguity rate Ti/j	20%	24%	20%	25%
Ambiguity rate Tj/i	27%	25%	21%	26%

The obtained results illustrate the important ambiguity between phonemes. This shows that, in acoustic level, it is difficult to distinguish between phonemes. To remedy this problem, it needs a linguistics study to integrate new parameters that permitto determine carefully the phonemes boundaries.

Ambiguity, in general, takes place between the neighbour phonemes or phonemes that have nearly the same pronunciation. In this context, there is an ambiguity between /T/ and /TT/, /S/ and /SS/ and /O/ and /U/. There is also another ambiguity between phonemes that are close in speech signal, for example between /A/ and /DD/ in number '*CRAD'* (three). This ambiguity is due to the interaction of the acoustics features.

6. Conclusion

The phonemes classification is the method that can classify a speech units based on the acoustics features. This classification can be used as a classifier for the hidden Markov model or neural network and it permits to improve the recognition rate and reduce the speech units' ambiguity. The Gaussian mixture is the most popular model used in classication. It is based on the three-dimensional presentation of data based on the vectors of the average and covariance matrices that model in a better way variation of speech signal. The obtained results show that there is an ambiguity between phonemes and there are no exact boundaries in speech signal. To resolve this problem, a special language study for each dialect must be made to take into account other characteristics of speech signal.

References

[1] Y Zhang, M Alder, R Togneri. *Using Gaussian Mixture Modeling for Phoneme Classification*. Centre for Intelligent Information Processing System Department of Electrical and Electronic Engineering The University of Western Australia. 2003.
[2] S Jamoussi. Méthodes statistiques pour la compréhension automatique de la parole. *Ecole doctorale IAEM Lorraine*. 2004.
[3] T Pellegrini, R Durée. Suivi de la voix parlée grâce au modèle de Markov caché. 1989.
[4] A Cornijeol, L Miclet. Apprentissage Artificielle: méthodes et concepts. 1988.
[5] SJ Young et PC Woodland. *The use of state tying in continuous speech recognition*. Proc, ESCA Eurospeech 1993. Berlin, Germany. 1993; 3: 2203-2206.
[6] H Bourlard, CJ Wellekens, H Ney. *Connected digit recognition using vector quantization*. Proc. IEEE Int. Conf. ASSP San Diego. CA. 1984.
[7] RM Gray. Vector quantization. *IEEE ASSP Mag.,* 1984; 1(2): 4-29.
[8] JL Gauvain, CH Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans, On Speech and Audio Processing*. 1994; 2(2): 291-298.
[9] D Jouvet, M Dautremont et Q. Gossart. Comparaison des multi modèles et des densités multi gaussiennes pour la reconnaissance de la parole par des modèles de Markov. *Actes des 20 èmes JEP*. 1994: 159-164.
[10] R André-Obrecht. A new statistical approach for the automatic segmentation of continuous speech signal. *IEEE Trans. On acoustics, speech, signal processing.* 1988; 36(1).
[11] F Jelinek. *Continous speech recognition by statistical methods.* Proceeding of IEEE. 1976; 64(4): 532-556.
[12] LA Liporace. *Maximum Likelihood estimation for multi-variant observation of Markov sources*. Proceeding IEEE trans IT. 1982; 28(5): 729-734.
[13] M Hwang, X Huang. *Sub phonetic modeling with Markov model.* Proceeding IEEE ICASSP-92, San Francisco, CA. 1992; 1: 33-36.