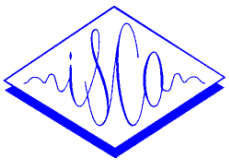

2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

COMPARISON OF TIME-DEPENDENT ACOUSTIC FEATURES FOR A SPEAKER-INDEPENDENT SPEECH RECOGNITION SYSTEM

D.Dubois

Centre National d'Etudes des Télécommunications
LAA/TSS/RCP route de Trégastel
22301 LANNION, FRANCE

ABSTRACT

Improvement of speech recognition performance was the goal to achieve with the experiments carried out in our laboratories. In this paper we shall attempt to show how we set out to maximize performance once the temporal variations of the coefficients were combined with the instantaneous vector as input to the HMM. Tests were conducted on several data bases (digits, isolated commands of a vocal server, and 2-digit numbers), recorded over the telephone. Following which, the concatenation of several frames, linear data reduction analysis techniques and linear regression data were tested.
Considerable improvement of the recognition performance was obtained by combining the first and second derivatives with the current frame over five adjacent frames. A recognition error rate of 0.7% was obtained. Normally, a 2.7% error rate is observed using exclusively cepstrum coefficients. This resulted in a 70% error rate reduction.
Keywords: HMM input coefficients; temporal variations; principal component analysis; discriminant analysis; linear regression.

1 INTRODUCTION

The speech recognition algorithm, which was developed at the CNET and which is based upon the Hidden Markov Model (HMM), uses an instantaneous vector of coefficients as its input. In order to enhance the performance of the algorithm, the introduction of the temporal variations of these coefficients was necessary. In fact, some phonemes are only discernable during a change in the acoustic spectra within a specified time period. The Furui study [1] demonstrates the important role of the transitions in the perception of phonemes and achieved significant improvements in the recognition system by combining both instantaneous and dynamic features. Taking these variations into account, Brown [2] proposed an association of several adjacent frames as input to the Markov system, and as such was able to use parameter reduction to decrease variations in the estimates of the acoustic event probabilities.
Calling upon the expertise cited in these studies, we examined different ways of introducing the temporal variation to the speech recognition system currently under study. First, an evaluation of the acoustic input was conducted, which resulted in the concatenation of two adjacent frames. Subsequently, a reduction of the input vector size was effectuated, along with the decorrelation of the parameters utilizing linear data reduction analysis techniques.
The primary linear method, the Principal Component Analysis (PCA), was utilized on several adjacent frames. The secondary linear method employed was the Discriminant Analysis (DA). Lastly, differential parameters and regression coefficients over 5 frames (80 ms of speech) were associated with the current frame. This last method was essential in the characterization of the formant slopes.
Following a brief description of the recognition system utilized and the data bases employed, each experiment will be described in detail.

2 SPEECH RECOGNITION SYSTEM

The speech recognition system utilized, [3], is based on the HMM associated with the Mel Frequency Cepstrum Coefficients (MFCC) which are computed every 16 ms. A version of this algorithm was implemented on the RDP50 PC board [4]. Experiments are currently under evaluation in terms of the potential for general public applications (voice interactive system and public voice mail systems [5, 6]). This algorithm is speaker-independent and is capable of recognizing both isolated and/or connected words.
Speech data is recorded over the telephone. The analog voice signal is digitized at 8kHz. The presence of speech data distortion is due to interference on the line.
After pre-emphasis, the signal is multiplied by a 32 ms Hanning window with 50% overlapping. From within the spectrum of FFT, 24 MEL filters compute 8 MFCC every 16 ms. Both the logarithm of total energy and its derivative are added after pre-emphasis. A vector of 10 coefficients, named edmfc, is subsequently obtained.
The recognition mechanism is based on HMM, and makes use of diagonal gaussian output distributions. Although different basic units can be utilized (phonemes, pseudo-diphones, allophones) regular n-state word models have been selected; n being equal to 13, 20 or 30 in the subsequent experiments. For each model, a loop on each state was authorized, capable of jumping to the next state, or two states away. Each transition, which achieves the same state, is associated with the same probability density function. The Viterbi algorithm is used for both the learning and recognition processes.

3 DATA BASES

Three data bases, composed of recordings conducted over the telephone lines, are currently available. The recordings were conducted in Lannion. Speakers, recorded over long distance lines from all over France, were used in order to obtain the maximum range of phonetic pronunciations. A manual verification of each recording was conducted. Only the correct elements were retained for use in the data base. Incorrect elements, sentences with speaking errors or truncated sentences, were discarded. For each database, the set of speakers were separated into two equal parts: one for training, the other for testing. The three data bases are broken down as follows:

Digit contained 10 digits (0 to 9) pronounced by 452 speakers, 226 speakers utilized for the learning phase, comprised of 2065 utterances and 226 speakers utilized for the testing phase, with 2080 utterances.

Tregor containing 36 isolated words, mostly used in a vocal server, offering local information regarding the Tregor area. 510 speakers participated in the recording of this vocabulary base. Hence, 255 voice recordings were utilized in the development of the learning phase, with 8354 utterances. The remaining 255 voice recordings were utilized for the testing phase, with 8402 utterances.

Number is a connected words data base composed of 2 digit numbers (00 to 99) recorded by 728 speakers. 364 voice recordings were used for the learning phase, with 6793 numbers, and 364 voice recordings were utilized for the testing phase, with 6708 numbers.

Most of the tests were carried out on the *Digit* data base for obvious reasons, due to the volume of calculation involved. The two other data bases, *Tregor* and *Number*, were used to confirm the results obtained on Digit.

4 EXPERIMENTS

Throughout the experiments conducted on the *Digit* data base, the standard reference used is the result cited in section 2, obtained with the parameter vector named *edmfc*. In this paper, "*n states*" (n=13, 20 or 30) is written to specify the number of states-per-word for the basic units used. The following table also cites the results obtained using a set of 9 coefficients (8 MFCC and energy) named *emfc*.

Table 1 : Error rate for standard sets of coefficients

Digit data base			
	13 states	20 states	30 states
emfc	5.9	4.8	3.5
edmfc	3.7	3.1	2.7

4.1 CONCATENATION OF TWO ADJACENT FRAMES

In the first experiment, two adjacent frames were combined into one vector composed of 20 parameters. The error rate obtained with a 13-state word model is 5.1%. A poor result of this nature may be due to the correlation of the input parameters that is not adapted to gaussian pdf's with diagonal covariance matrices. In order to reduce the input vector size, and more specifically, decorrelate the parameters, linear data analysis techniques were used.

4.2 LINEAR TRANSFORMATION METHODS

A reduction in the number of parameters was the ultimate aim, however, at the same time, our parallel aim was to safeguard as much information as possible. In order to do so, a transformation was provoked which reduced the vector of the parameters to a smaller dimension.
The first linear transformation method used is the Principal Component Analysis(PCA) on several adjacent frames, ranging from 1 to 5.

4.2.1 Principal Component Analysis

The covariance matrix of the eigenvector is calculated from the base of all the learning data. The new vector is obtained by the projection, onto each principal component, of an initial vector built from several adjacent frames.
The PCA was performed with one, two, three and five adjacent frames. The recognition tests were conducted utilizing a number of parameters varying between 10 and 25 (post PCA) and a 13-state word model. The best recognition error rate, obtained on the *digit* data base, was 3.7% with 20 PCA parameters (4.7% with 10 PCA parameters). When compared with the results obtained with the 10 original parameters *edmfc* (see Table 1), this is still not satisfactory. A possible explanation is that the variance due to the differences between the speakers and the telephone lines is greater than the variance between the actual words, which required the search for a more discriminating method.

4.2.2 Discriminant Analysis

The discriminant analysis (DA) method was used as the secondary linear transformation method. This method maximized the between-class variance to within-class variance ratio. In order to effectuate this test, the classification selected was that of phones, and the markers utilized were those which issued from the Markov model automatic labelling. The basic unit is a 13-state word. The DA was constructed using 1 to 3 frames, namely 10-to-30 coefficients; the best error rates (4.5% for 10, 4.4% for 15, 3.7% for 20), are still inferior to those obtained using the *edmfc* coefficients.
The results obtained with this kind of analysis were analogous to those obtained with the PCA method and were less effective than the instantaneous frame. One of the problems originated from the initialization phase, there being no manual verification of the markers present. Another problem derived from the fact that the sentences present in the three databases contained background noise which perturbed the covariance matrix. As a result, a Speech/Noise detection was introduced prior to calculating the covariance matrix, but the results obtained were no better.
Our work was therefore oriented towards techniques that included, at every frame, the current parameters and other parameters calculated over several frames.

4.3 DIFFERENTIATED CEPSTRUM COEFFICIENTS

The differential parameters initially tested were the differentiated coefficients, named *delta*, obtained by simply associating the differences calculated between the following frame and the previous frame (48 ms of speech) with the current frame. The resulting 18 coefficients vector consist of 9 coefficients *emfc* for the current frame, and the 9 differenciated coefficients. Table 2 gives the error rate for the 13, 20 and 30 states word models.

Table 2 : Error rate with Differenciated Coefficients

	13 states	20 states	30 states
delta	2.4	2.2	1.7

These results can be compared with those of table 1 obtained with *edmfc*. The speech recognition algorithm error rate was reduced by about 33%.

4.4 LINEAR REGRESSION COEFFICIENTS

In this experiment linear regression coefficients were used to extract temporal changes in the spectra. We simply associated to the current frame the slope of the linear regression of the coefficients over 3 or 5 frames (48 ms or 80 ms of speech). The resulting 18 coefficients vector consisted of 9 *emfc* coefficients and 9 slopes coefficients. We noted *emfcdi* (i equal 3 or 5 represents the number of frames) the sets of parameters obtained.

Table 3 : Error rate with Linear regression Coefficients

	13 states	20 states	30 states
emfcd3	2.8	2.1	1.9
emfcd5	2.2	1.5	1.2

Certain observations may be made concerning these experiments. First of all, significant enhancement was obtained by joining the difference of energy between two frames (*emfc* versus *edmfc* Table 1). The linear regression over 3 frames produces results similiar to that of the delta coefficients; results which were also observed by K.F.Lee [7]. However, the linear regression over 5 frames decreased the error rate for the 30-state word model, by about 36% compared with *emfcd3*. This set of parameters was retained for later experiments.

A complementary test was carried out from *emfcd5*: we replaced the first 9 coefficients (8 MFCC and energy) of the current frame by mean over 5 frames. The 9 slopes of the linear regression over 5 frames was retained. This set of parameters was tested using a 13-state word model and the error rate obtained was 6% instead of 2.2% with *emfcd5*. So, by using the mean value of the coefficients over a large window instead of instantaneous coefficients the results were unsatisfactory. This fact can explain the inefficiency of data reduction methods.

4.5 VARYING NUMBER OF COEFFICIENTS

The set of parameters providing the smallest error rate with the digit data base (1.2%) is *emfcd5*. 18 coefficients are necessary in order to obtain this result. Our aim was to decrease the number of coefficients without increasing the error rate. The interesting aspect of this experiment is the reduction of calculation time. The *emfcd5* coefficients are arranged in the following order: E, C1 to C8, ∂E, ∂C1 to ∂C8 with Ci for MFCC and ∂Ci for MFCC derivative. 6 to 18 coefficients *emfcd5* were tested with a 13-state word HMM basic unit.

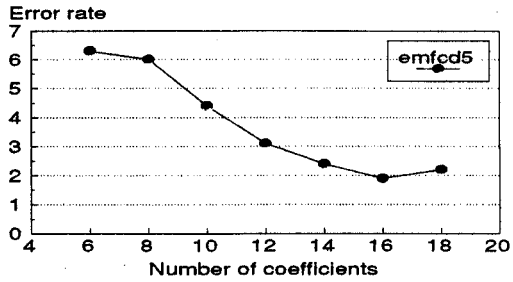

Figure 1 : Influence of the number of coefficients

One important fact that can be deduced from this figure is that after a marked improvement at the beginning of the curve, the enhancement is not significant at the end of the curve (from about 14 coefficients). To verify the influence of the number of coefficients and the improvements due to the MFCC derivative associated to the current frame, we used the three data bases (*Digit*, *Tregor*, and *Number* data bases), three HMM word basic unit: 13 states, 20 states, 30 states and four sets of parameters: *edmfc*, *emfcd5/14*, *emfcd5/16*, *emfcd5/18* (*emfcd5/n* meaning n first coefficients of *emfcd5*).

4.5.1 Digit data base

Table 4 : Error rate on Digit data base

	13 states	20 states	30 states
edmfc	3.7	3.1	2.7
emfcd5/14	2.6	1.8	1.5
emfcd5/16	1.9	1.6	1.3
emfcd5/18	2.2	1.5	1.2

These results confirm that above a certain number of coefficients the gain gets smaller. The confidence interval is 0.4%, so the difference in the error rate is insignificant.
Complementary tests were performed with the two other data bases *Tregor* and *Number*.

4.5.2 Tregor data base

Table 5 : Error rate on Tregor data base

	13 states	20 states	30 states
edmfc	4.6	4.0	2.6
emfcd5/14	1.8	1.3	0.9
emfcd5/16	1.8	1.4	0.9
emfcd5/18	2.0	1.2	0.9

The information gathered from the Digit data base is still valid for this data base both as regards the HMM basic unit and also the number of coefficients.

4.5.3 Number data base

All the tests realized to date were conducted on isolated words; however, the goal to be achieved was the validation of results on connected words.

Table 6 : Error rate on Number data base

	13 states	20 states	30 states
edmfc	17.0	11.9	9.6
emfcd5/14	10.2	6.9	5.5
emfcd5/16	10.4	6.4	5.5
emfcd5/18	10.6	6.4	5.5

As in the preceding cases the conclusions of these experiments were extremely promising. There was a considerable reduction in error rates when the first derivative adjoining the current frame was used. However, the reduction was less when used over a certain number of coefficients.

4.6 CONTRIBUTION OF THE SECOND DERIVATIVE

Some tests were carried out on an extended *digit* data base containing 3563 digits for the learning phase and 3614 digits for tests pronounced by 775 speakers. The sets of parameters tested are named *emfcd5/l.m.n.* with "***l***" first coefficients of *emfc*, "***m***" first coefficients of first derivative and "***n***" first coefficients of second derivative. The complete set of parameters, named *emfcd5/9.9.9.*, are composed of 9 *emfc*, 9 first derivatives and 9 second derivatives computed over 5 frames. All the attempts were made with a n-state word basic unit, (noted n st in Table 7), n varying between 13 and 41. The results obtained can be easily compared with those obtained with the "old" *Digit* data base.

Table 7 : Error rate with second order variances

	13 st.	17 st.	21 st.	29 st.	41 st.
emfcd5/9.1.0.	3.5	3.2	3.1	2.4	2.2
emfcd5/9.9.0.	2.0	1.6	1.3	1.0	1.0
emfcd5/7.7.4.	1.6	1.4	1.3	0.9	0.9
emfcd5/9.9.9.	1.5	1.0	0.9	0.8	0.7

Firstly, it is important to observe that the error rate decreases when the number of states increases. With the standard reference set, *emfcd5/9.1.0*, which is the same as *edmfc*, the error rate is reduced by 37% between 13 and 41 states. In the same way, the error rate decreases when we adjoin the first and second derivatives. The *emfcd5/9.9.0* set results in greater enhancement ranging from 42 up to 58%. This depends the models. The improvement achieved by the second derivative, the *emfcd5/9.9.9* set, is also significant, the error rate is still reduced by about 15% (from 57% up to 70% with the reference set). The best error rate obtained is 0.7%.

On the level of equal complexity, which can be measured with the total number of model parameters, we shall check that it is preferable to use the first derivatives than to double the number of states. For instance, the model with 13 states and the *emfcd5/9.9.0* set have the same complexity as the model with 29 states and the *emfcd5/9.1.0* set. However, the result is better (2.0% versus 2.4%). Using the model including 17 or 21 states and the *emfcd5/9.9.0* set, the results achieved were better (1.6% or 1.3% versus 2.2%) than when the model with 41 states and the *emfcd5/9.1.0* set were used. The complexity is the same in both cases.

5 CONCLUSION

Because of the important role played by temporal changes in the spectra, we investigated several ways of enhancing our speaker-independent speech recognition system.

Our initial tests using automatic methods to extract pertinent parameters were unsuccessful. The principal component analysis and the linear discriminant analysis, this last method being more phonetically oriented data analysis, were therefore chosen. We were looking for techniques that would take into account time-dependent information calculated over several frames, in addition to the current frame.

Utilizing an 18 parameter vector, composed of differenciated cepstrum coefficients over 3 frames and the current frame, we reduced the word error rate by 37% with a 30-state word basic unit.

Utilizing an 18 parameter vector, composed of the current frame and linear regression coefficients over 5 frames, as well as a 30-state word basic unit, the error rate is decreased by 55%. A final attempt using the second derivatives as a supplement to the other 18 coefficients gives a further 15% reduction in the error rate. This represents an overall 70% reduction with regard to the standard error rate given by *emfcd5/9.1.0.*

In conclusion we have found that by joining first derivative coefficients it is easier to achieve enhancements than by adding states. The error rate can then be reduced further by joining the second derivatives and some states to the word basic unit. The minimum error rate obtained using the extended digit data base is 0.7%.

REFERENCES

[1] S.FURUI: *Speaker-Independent Isolated Word Recognition Using Dynamic features of Speech spectrum*, IEEE Trans. ASSP, vol.ASSP-34, pp.52-59, Feb.1986.

[2] P.F.BROWN: *The acoustic-Modeling Problem in Automatic Speech recognition*, IBM Thomas J.Watson Research center, Distribution Services 73-F11, Post Office Box 218, Yorktown Heights, NY 10598. 1988.

[3] D.JOUVET, J.MONNE, D.DUBOIS: *A new network-based speaker independent connected word recognition system*, IEEE ICASSP 1986, Tokyo.

[4] J.P.TUBACH, C.GAGNOULET, J.L.GAUVAIN: *Advances in speech recognition products from France*, Speech Technology Conference, New-York, Avril 1989.

[5] C.GAGNOULET: *Speech recognition over the telephone:Experiments in France*, Voice systems worldwide'90 Londres.

[6] C.GAGNOULET, D.JOUVET: *Développements récents en reconnaissance de la parole*, L'Echo des Recherches N° 135, 1989.

[7] K.F.LEE: *Large-Vocabulary Speaker-Independ- ent Continuous Speech Recognition: The SPHINX System*, PhD Carnegie Mellon University, CMU-CS-88-148, April 1988.