Speaker Recognition Using Dynamic Features and The Null Space Combination of Classifiers in The Sub-band Domain

Eduardo Esteves Vale, Abraham Alcaim
CETUC-PUC/Rio
Rio de Janeiro-RJ-Brazil

Rosângela Coelho
Instituto Militar de Engenharia – IME
Rio de Janeiro-RJ-Brazil

Abstract—**This paper investigates the inclusion of dynamic features in the input vectors used by the multiple classification scheme which employ the null space to combine the likelihoods. Speaker identification experiments were performed considering four ambient noise and also different mismatched conditions. The results show that this strategy can contribute to increase the recognition accuracy.**

Keywords—**Delta features, delta-delta features, multiple subband-classifiers, speaker recognition.**

I. INTRODUCTION

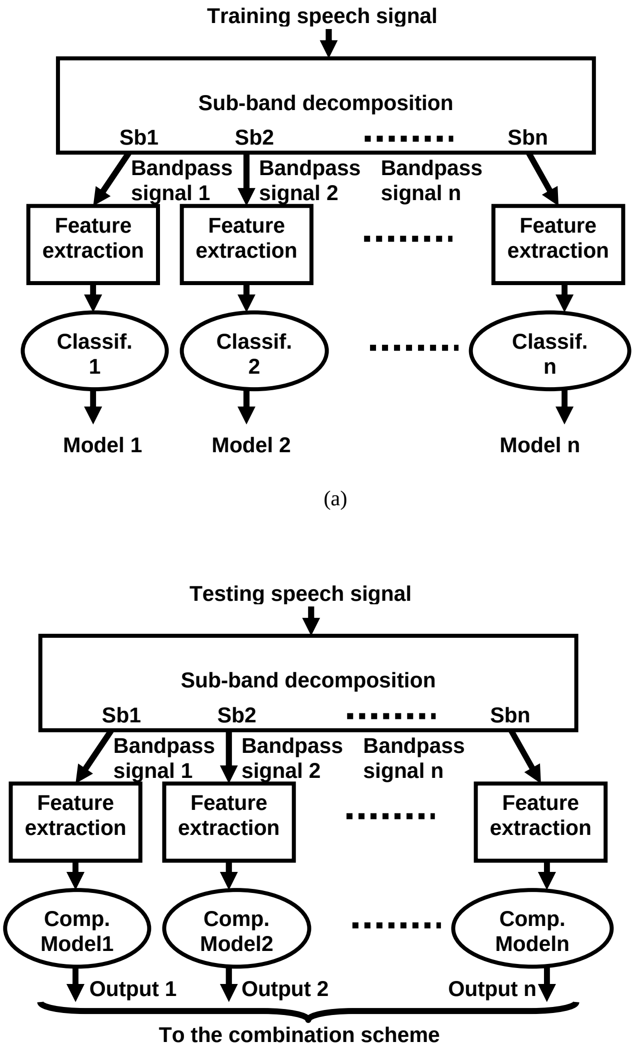
The speaker recognition systems are widely employed in many situations nowadays. They can be used in criminal investigations, security systems and other applications. However, in most cases, the performance of the recognition is severely affected by environmental noises that can often be present in the speech signals. It has been shown [1]-[4] that strategies employing the sub-band processing of the signal contributes to overcome the effect of the noise. In these techniques, the main goal is to better use those bandpass signals which are more important for the recognition. The multiple sub-band classifier systems explore this advantage that the sub-band decomposition can provide.

Interesting improvements in the recognition accuracy have been achieved by using combination strategies of classifiers outputs [3], [4]; and, in some situations, by using multicondition training in the sub-band domain [5]. The combination technique is used to favor those sub-band outputs which are more related to the speaker recognition. This is because the noise contribution, and the identity information [6] are nonlinear distributed [7] among the bandpass signals provided by the decomposition. Fig. 1 shows a speaker recognition system using multiple sub-band classifiers. As can be seen, the speech signal is decomposed into *n* sub-bands. The feature extraction is performed in each bandpass signal and used as input for a classifier in that band. Each classifier, in the training phase, generates a probabilistic sub-band model of a speaker. During the testing phase (Fig. 1 (b)), the sub-band features are compared with the sub-band speaker model. The likelihood resulted of this comparison is joined to the other ones in a combination scheme, which produces the joint response for a modeled speaker.

The literature presents some classifiers' outputs combination strategies. In a first approach [1], the classifiers' output were combined by the sum of these outputs. In another approach [3], the outputs were combined with non-uniform weights calculated by the total energy of each bandpass training signal. The same weights are used for every recognition test. The employment of non-uniform weights can better represent the nonlinear distribution of the identity information. This combination rule provides low computational cost and spend small memory space. However the improvements in performance are small for tests with white noise. Better results were obtained for the case of colored noise (i.e. non-white). Another strategy [4] consisted in obtain a collection of weights provided by the null space calculated from the total energy of each bandpass training signal. The main advantage is that the weights can be changed during the tests. Improvements were obtained for colored and white noises. According to the proposal presented in [5], the use of the null space [8] combination rule together with multicondition training with white noise compensates the effect of the noises. The improvements were obtained only when the test signals were corrupted by car and white noises. It is necessary to better choose the type of noise used to train the recognition system in order to improve the performance of the method for other types of noises. However, this is very difficult since there are many types of noises with a large variety of behavior in frequency.

This work examines the inclusion of dynamic features in the input MFCC (Mel Frequency Cepstral Coefficients) [9] feature vectors of the sub-band GMM (Gaussian Mixture Model) [10] classifiers, in order to observe how the performance of the recognition system using the null space can be improved. Several experimental results of text-independent speaker identification [11] using a large speech material and different types of environmental noises are shown in order to demonstrate the performance of the recognition techniques.

The rest of the paper is organized as follows. Section II describes the MFCC features used in this paper. Section III describes the dynamic features (delta and delta-delta). Section IV presents the GMM classifier. Section V describes the combination scheme that employs the null space. The experiments, results and discussions are presented in Section VI. Section VII concludes the paper.



(a)

(b)

Figure 1. Speaker recognition system using multiple classifiers in the sub-band domain: (a) Training to obtain the *n* sub-band Models of a speaker; (b) testing the *n* sub-band Models of a speaker to obtain the combined response.

II. Mel Frequency Cepstral Coefficients

The MFCC are static features extracted by a filter bank which models human perception of the frequency content of sounds [9]. This perception follows a subjectively defined nonlinear scale called the “mel” scale, defined as,

$$f_{mel}=2595\log_{10}(1+f\,/\,700) \qquad (1)$$

where f is the actual frequency in Hz. The features can be calculated as follows: first the DFT (Discrete Fourier Transform) is applied to a frame of speech. Next, triangular filter banks, that are linearly spaced in the mel scale, are imposed on the spectrum. Finally, DCT (Discrete Cosine Transform) is taken on the log filter bank energies.

III. Dynamic Features

The delta and delta-delta features, also known as dynamic features [10], complement the instantaneous or static information obtained by the MFCC. The delta-MFCC feature vector represents the time derivative of the MFCC features. The dynamic features represent spectral changes over time. In addition, these features can remove time-invariant spectral information. It can be expressed by

$$\Delta f_k[i]=f_{k+M}[i]-f_{k-M}[i], \qquad (2)$$

were $f_k[i]$ denotes the *i*th feature in the *k*th time frame, *M* is typically 2-3 frames, and $\Delta f_k[i]$ is the delta parameter of the *i*th feature. The delta-delta feature can be obtained by the delta feature using the same principle. In the experiments presented in this paper, the dynamic features employed are the delta and the delta-delta obtained from the MFCC features, with the purpose of removing the local time-invariant information of noise signals.

IV. Gaussian Mixture Model

The GMM algorithm [10] models a distribution by a mixture (weighted sum) of M Gaussian probability densities. This mixture can be expressed as,

$$p(\vec{x}|\lambda)=\sum_{i=1}^{M} p_i b_i(\vec{x}) \qquad (3)$$

where p_i are the weights, λ represents the mixture model, $\vec{x}$ is a random vector of dimension D, and the $b_i(\vec{x})$ are the density components of the form

$$b_i(\vec{x})=(1/(2\pi)^{D/2}|\Sigma_i|^{1/2})\exp\{-(1/2)(\vec{x}-\vec{\mu}_i)'\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)\}, \qquad (4)$$

where Σ_i is the covariance matrix, $\vec{\mu}_i$ is the mean vector, and $(\vec{x}-\vec{\mu}_i)'$ is the transpose of $(\vec{x}-\vec{\mu}_i)$. The parameters of the mixture of densities (mean vectors, covariance matrices, and weights), that represent the model, are estimated iteratively by the expectation-maximization (EM) algorithm [12]. In the test phase of the model, a likelihood is obtained by introducing the testing input feature vector into the mixture density function (3) using the set of model parameters. The resulting log-likelihood for an utterance can be obtained by the sum of the logarithm of the likelihoods calculated using each testing feature vector. In the decision of the identification scheme, this can be applied as

$$\hat{S} = \arg \max_{1 \le k \le S} \sum_{t=1}^{T} \log p(\vec{x}_t|\lambda_k) \qquad (5)$$

where S is the number of speakers and T is the number of feature vectors.

V. Recognition Using The Null Space

The linear algebra states that the null space of an *m* by *n* matrix A is a vector space formed by all the solutions of the homogeneous system, Ax=0, where x represents the *n* dimensional solution vector [8]. The computation of a basis for the null space is an operation which provides a set of linearly independent solutions that is closed under addition and scalar multiplication. This means that the summation of these vectors or any multiple of them is also a solution for Ax=0. The main idea of this scheme is to use the total energy of each bandpass training signal, as the column elements of A. Hence, A is defined from important speaker information in different sub-bands. Note that, in this case, A is a 1 by *n* matrix, where *n* is the number of sub-bands. Therefore, every column of A represents a speaker sub-band energy. The computation of a basis for the null space of A provides a set of *n*-1 solution vectors $\{\times_1, \ldots, \times_{n-1}\}$ which retain the speaker dependent information. These vectors provide *n*-1 ways to represent the speaker dependent information without redundancies, since these solutions are linearly independent. The elements of the solution vectors are used as weights to be applied to the sub-band classifiers outputs, associated to the modeled speaker. During the test of a speaker, all the reference vectors related to that speaker are tested in order to find which of them better contributes for the recognition task. The one which provides the highest likelihood is chosen.

VI. Experimental Results

This section presents experimental results of text-independent speaker identification obtained in order to show the behavior of the scheme using the dynamic features, when compared to the other methods. In this experiment, the speech signals were corrupted by colored noises (Factory1, Babble and Volvo) from the NOISEX-92 database [13] and by White Gaussian noise generated by a Matlab tool. It was used 49 speakers (male) and their corresponding clean speech signals (sampled at 8 kHz) obtained from sessions 1 to 5 of the KING database [14]. The experiments of this article were conducted using a subset of this speech database, which is a collection of conversational speech from male speakers. For each speaker there are 10 conversations recorded during 10 separate sessions. The speech from a session was locally recorded from a high-quality microphone and was transmitted over a long distance telephone link, providing a high-quality (clean) version and a telephone quality version of the speech. The experiments presented in this paper, use only the clean version of this database. The sub-band classifier chosen to perform the experiments is the GMM since it is a powerful statistical tool extensively used for many speaker recognition applications. The clean signals from sessions 1, 2 and 3, without silence, were used [15] to train the GMM classifiers with 90 seconds of speech. The signals from the remaining two sessions, composed by four segments of 15 seconds (for every speaker and without silence), were corrupted by noises at 10 dB and 15 dB of SNR and used for test. In addition, recognition with four segments of 5 seconds was performed in order to show the behavior of the schemes. It was used 20 MFCC parameters (with their 20 delta and/or delta-delta features appended, when dynamic features are included), extracted in frames of 20 ms of speech signals (Hamming windowed and overlapping by 50%). The techniques presented in [1],[4] and the scheme with dynamic features used four sub-bands (Sbs) produced by mel-spaced 6th order Butterworth filters. The GMM classifiers used 32 gaussians (M=32 probability densities) to obtain the speaker model.

The experimental results expressed in terms of recognition rate are presented in Tables I to IV. This measure is given by RR (%) = (number of correct identification / number of tests) × 100 %. The identification performance obtained by using only one GMM (without sub-band decomposition) is 96.43% for test in 15s of speech without noises. This performance severely drops in speech corrupted by environmental noises. In the Tables I to IV, the multiple sub-band classifier approach which the combination technique consists in summing the outputs is represented as Sum [1], the one that employs the null space is represented as Null space [4], and the proposed, which uses the null space and the dynamic features, is represented as *N. space and Delta*, *N. space and Delta-delta* (when delta or delta-delta are appended in the MFCC feature vector) and *N. s. and Delta, Delta-delta* (when both delta and delta-delta are appended in the same MFCC feature vector). Table I presents the recognition rate for tests using utterances of 15 seconds in 15 dB of SNR.

Table I. RR(%) in 15s and with noises in 15 dB of SNR

	Factory1	Babble	Volvo	White
Sum (4Sbs) [1]	66.33	72.45	67.86	30.10
Null space (4Sbs) [4]	75.51	81.12	76.53	40.82
N. space and Delta (4Sbs)	75.00	80.61	78.06	39.29
N. space and Delta-delta(4Sbs)	74.49	80.57	79.08	37.76
N. s. and Delta, Delta-delta (4Sbs)	78.06	83.16	78.06	40.82
1GMM	70.41	76.35	73.47	27.51

The best result of 78.06% is obtained for the null space scheme with dynamic features, when the test speech is corrupted by Factory noise. When the test speech is corrupted by Babble noise, the highest performance of 83.16% is also obtained for the null space scheme with dynamic features. For the Volvo noise, the best result of 79.08% is obtained for the proposed scheme with delta-delta features. Finally, for the case of white noise, the highest result of 40.82% is obtained for the null space techniques with and without dynamic features.

Table II presents the recognition rate for tests using 5 seconds of speech in 15 dB of SNR.

TABLE II. RR(%) IN 5S AND WITH NOISES IN 15 DB OF SNR

	Factory1	Babble	Volvo	White
Sum (4Sbs) [1]	52.55	59.69	57.65	26.02
Null space (4Sbs) [4]	60.71	68.37	65.31	35.71
N. space and Delta (4Sbs)	64.80	68.35	64.29	36.73
N. space and Delta-delta(4Sbs)	65.31	66.84	62.76	32.65
N. s. and Delta, Delta-delta (4Sbs)	65.31	69.90	66.33	33.67
1GMM	63.78	65.19	62.22	26.53

From this Table, It can be seen that the highest result of 65.31% is obtained for the proposed schemes using dynamic features, when the test speech is corrupted by Factory noise. When the test speech is corrupted by Babble noise, the highest performance of 69.90% is also obtained for the null space scheme with dynamic features. For the Volvo noise, the best result of 66.33% is again obtained for the null space scheme with dynamic features. Finally, for the case of white noise, the highest result of 36.73% is obtained for the proposed technique using the delta features.

Table III presents the recognition rate for tests using 15 seconds of speech in 10 dB of SNR.

TABLE III. RR(%) IN 15S AND WITH NOISES IN 10 DB OF SNR

	Factory1	Babble	Volvo	White
Sum (4Sbs) [1]	34.69	69.39	43.88	17.35
Null space (4Sbs) [4]	48.98	80.10	58.16	25.51
N. space and Delta (4Sbs)	51.53	80.14	59.18	26.50
N. space and Delta-delta(4Sbs)	57.14	80.10	66.33	24.49
N. s. and Delta, Delta-delta (4Sbs)	55.61	80.10	68.88	25.00
1GMM	34.18	63.27	33.67	10.20

In this Table the highest recognition rate of 57.14% is obtained for the proposed scheme using delta-delta features, when the test speech is corrupted by Factory noise. When the test speech is corrupted by Babble noise, the highest performance of 80.14% is obtained for the proposed scheme using delta features. For the Volvo noise, the best result of 68.88% is obtained for the proposed scheme with delta and delta-delta features. Finally, for the case of white noise, the highest result of 26.50% is obtained for the proposed technique using the delta features.

Table IV presents the recognition rate for tests using 5 seconds of speech in 10 dB of SNR.

TABLE IV. RR(%) IN 5S AND WITH NOISES IN 10 DB OF SNR

	Factory1	Babble	Volvo	White
Sum (4Sbs) [1]	33.80	47.45	41.02	17.23
Null space (4Sbs) [4]	40.51	57.14	58.03	22.45
N. space and Delta (4Sbs)	42.06	57.16	57.14	22.96
N. space and Delta-delta(4Sbs)	52.04	59.69	57.10	21.94
N. s. and Delta, Delta-delta (4Sbs)	53.06	59.18	58.16	22.45
1GMM	33.76	55.24	33.06	9.69

Table IV shows that the highest result of 53.06% is obtained for the proposed scheme using dynamic features, when the test speech is corrupted by Factory noise. When the test speech is corrupted by Babble noise, the highest performance of 59.69% is obtained for the proposed scheme using delta-delta features. For the Volvo noise, the best result of 58.16% is also obtained for the null space scheme with dynamic features. Finally, for the case of white noise, the highest result of 22.96% is obtained for the proposed technique using the delta features.

Tables I to IV showed that for all cases the proposed technique which uses the dynamic features achieves the highest performance. Particularly for 15dB, the best performance in most cases is obtained when both delta and delta-delta are presented in the same feature vector. However, it’s not true for 10dB.

Note that in some evaluations the dynamic features do not improve the recognition, as for example in [16], [17]. However, when the test speech signal is less affected by noise, the inclusion of several dynamic features can contribute to increase the performance of the system due to the additional dynamic information. In clean environment the delta and delta-delta features are usually more adequate for text-dependent speaker identification and for applications that require the reduction of channel mismatch [10]. Moreover it can be seen that when the test speech is very affected by noise (SNR=10dB), the contribution due to the MFCC becomes very small. However, the contribution due to the inclusion of dynamic features tends to increase the performance of the recognition. The noise penalizes more the contribution of the MFCC feature (which is not robust) rather than that of the dynamic features.

The results provided by the recognition system without sub-bands are expressed in the row of the 1 GMM and are used only for a reference.

VII. CONCLUSIONS

This paper proposed the inclusion of dynamic features (delta and delta-delta) in the input vectors of the sub-band classifiers, in order to observe how the performance of the recognition techniques using the null space can be improved in text-independent speaker recognition in noisy environments. We have performed experiments with the dynamic features extracted in sub-bands of frequency applied to the multiple sub-band classifier system which uses the null space. The results obtained show that the inclusion of dynamic features is capable to increase the performance of the recognition.

Several experimental results of text-independent speaker identification using a large speech material and different types of environmental noises have been presented in order to demonstrate the performance of the recognition systems.

REFERENCES

[1] J. E. Higgins, R. I. Damper and T. J. Dodd, "Information fusion for subband-HMM speaker recognition," Proc. IEEE Int. Joint Conf. on Neural Networks, vol. 2, pp. 1504-1509, Washington DC, July 2001.

[2] Z. Sakka, A. Kachouri, A. Mezghani e M. Samet, "A new method for speech denoising and speaker verification using subband architecture," First Int. Symp. Control, Communications and Signal Processing, pp. 37-40, Hammamet, March 2004.

[3] E. E. Vale, A. A. Cunha and A. Alcaim, "Robust text-independent speaker identification using multiple subband-classifiers in colored noise environment," Int. Conf. on Systems, Signals and Image Processing, pp. 275-277, Bratislava, June 2008.

[4] E. E. Vale and A. Alcaim, "Adaptive weighting of subband-classifier responses for robust text-independent speaker recognition," Electronics Letters, vol. 44, issue 21, pp. 1280-1282, October 2008.

[5] E. E. Vale and A. Alcaim, "Combinação das respostas de classificadores em sub-bandas para reconhecimento de locutor usando espaço nulo e treinamento com múltiplas condições," XVII Simp. Brasileiro de Telecomunicações, pp. 4 pp-CD-ROM, Blumenau, October 2009.

[6] S. Wenndt and S. Shamsunder, "Bispectrum features for robust speaker identification," IEEE Int. Conf. on Acoustics, Speech and Signal Processing, vol. 2, pp. 1095-1098, Munich, April 1997**.**

[7] H. Guangrui and W. Xiaodong, "Improved robust speaker identification in noise using auditory properties," Proc. Int. Symp. Intelligent Multimedia, Video and Speech Processing, pp. 17-19, Hong Kong, May 2001.

[8] G. Strang, Linear Algebra and Its Applications,. Thomson Learning, 1988.

[9] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. on Acoustics, Speech and Signal Processing, vol. ASSP-28, no. 4, pp. 357-366, August 1980.

[10] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. on Speech and Audio Processing, vol. 3, no. 2, pp. 72-83, January 1995.

[11] H. Gish and M. Schmidt, "Text-independent speaker identification," IEEE Signal Processing Magazine, vol. 11, issue 4, pp. 18-32, October 1994.

[12] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," JournalRoyalStat. Soc., vol. 39, pp. 1-38, 1977.

[13] A. Varga, H. J. M. Steeneken, M. Tomlinson and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," Technical Report, DRA Speech Research Unit, 1992.

[14] J. Godfrey, D. Graff and A. Martin, "Public databases for speaker recognition and verification," Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 39-42, Switzerland, April 1994.

[15] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4, pp. 639-643, October 1994.

[16] Y. Shao and D. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1589-1592, Nevada, March-April 2008.

[17] L. Liu, J. He and G. Palm, "Signal modeling for speaker identification," IEEE Int. Conf. on Acoustics, Speech and Signal Processing, vol. 2, pp. 665-668, Atlanta, May 1996.