Multitaper MFCC Features for Acoustic Stress Recognition from Speech

Salsabil Besbes
University of Tunis El Manar
National School of Engineers of Tunis
Signal, Image and Information Technology laboratory
BP. 37 Le Belvdre, 1002, Tunis, Tunisia

Zied Lachiri
University of Tunis El Manar
National School of Engineers of Tunis
BP. 37 Le Belvdre, 1002, Tunis, Tunisia

***Abstract*—Ameliorating the performances of speech recognition system is a challenging problem interesting recent researchers. In this paper, we compare two extraction methods of Mel Frequency Cepstral Coefficients used to represent stressed speech utterances in order to obtain best performances. The first method known as traditional is based on single window (taper) generally the Hamming window and the second one is a novel technique developed with multitapers instead of a single taper. The extracted features are then classified using the multiclass Support Vector Machines. Experimental results on the SUSAS database have shown that the multitaper MFCC features outperform the conventional MFCCs.**

***Keywords*—Mel Frequency Cepstral Coefficients (MFCC); Multitapering; Multiclass SVM; Stress recognition**

I. Introduction

Automatic stress speech recognition have been used in several applications such as human to machines communications, craft voice communications, medicine and psychology. Stress refers to human response to different factors such as workload task, environmental condition and health. Stress has an impact in the performance of a person in his daily life. It affects the brain, the muscles, the eyes, the cardiovascular system and especially the speech production system.

Stress recognition systems are composed of two important steps which are feature extraction and feature classification. In literature, different classifiers have been used including Hidden Markov Model(HMM) [1] , Artificial Neural Network systems(ANN) [2], Gaussian Mixture Model(GMM) [3] and Support Vector Machines (SVM) [4].

In last years, extracting the most suitable features set for stressed speech recognition has been an important subject in many researches. Feature extraction aims to obtain a compact representation of speech signals. Studies have proposed different acoustic features to represent the speech under stress signals. These features are essentially Pitch, energy, [5], Linear Predictive Cepstral Coefficients(LPCC) [6] and Mel Frequency Cepstral Coefficients (MFCC). Different features have been extracted in [7] including pitch, energy, formants and MFCC from the stressed speech.

The MFCC features are the most common used features in speech processing because they are based on human auditory system. Usually, MFCCs are computed using a windowed periodogram via the Discrete Fourier Transform (DFT) [8]. It has been demonstrated that the spectrum estimate obtained has a high variance despite it low bias. A solution was proposed to reduce the spectral variance using multitaper spectrum estimate instead of the single windowed periodogram [9], [10]. In order to have a low variance spectrum estimate, the multitaper method applied a set of orthogonal tapers to the speech signal and an average sum of the sub-spectra are then calculated.

The multitaper approach have been used in several domains including geophysical applications [11], speaker verification [12], [13] and emotion recognition [14], [15] and it has been shown to improve the performance and robustness of different systems. However, this method has not been used in stress speech recognition applications. So, the aim of this work is to investigate multitaper MFCC features (MMFCC) in order to improve the performances of the stressed speech recognition system. We are also interested to compare different methods of multitapering including Thomson method, Multipeak method and SWCE (Sinusoidal Weighted Cepstrum Estimator) method.

This paper is structured as follows: Section II present the stressed speech recognition system proposed in this work. Section III presents the process of multitaper MFCC extraction and Section IV describes the multitaper spectrum estimate method. Section V deals with multiclass Support Vector Machines approaches. Results and experiments are given in Section VI. Finally, conclusion is presented in Section VII.

II. System Framework

The system proposed for stress speech recognition is illustrated in figure 1. First, we extract MFCCs and MMFCCs from the stressed speech signals of the SUSAS database. These features are then divided into training and test sets. Second, the classification is realized based on multiclass SVM methods which are One-Versus-Rest(1vR), One-Versus-One(1v1) and Directed Acyclic Graph (DAG). The performance of the system is evaluated with accuracy rate using the test set.

III. Multitaper MFCCs Extraction

In this section, we describe the extraction procedure of the multitaper MFCC features in order to compare it to the traditional extraction technique of MFCC features [8].
The traditional MFCC coefficients can be obtained following the same steps described for MMFCCs and using a special case

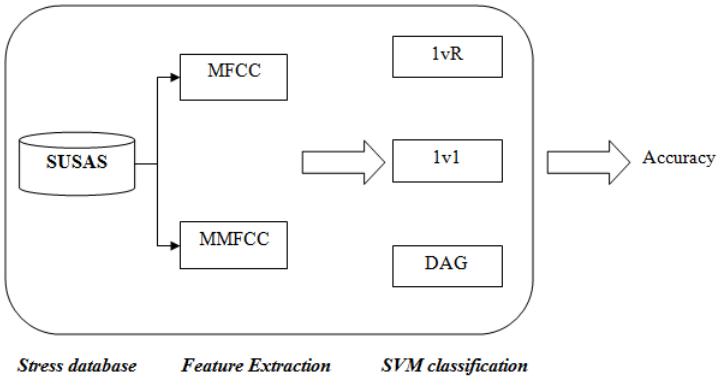

Fig. 1: System framework

of multitaper spectrum estimate which leads to a Hamming windowed spectrum.

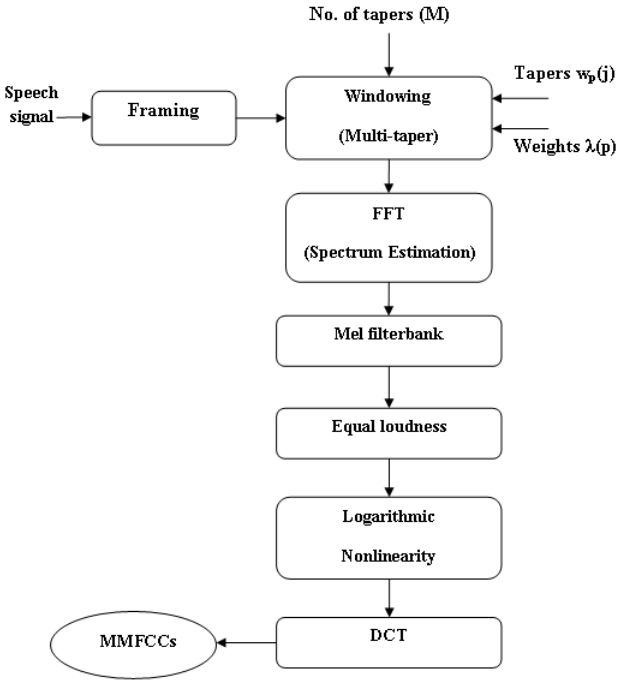

Fig. 2: Extraction procedure of multitaper MFCC

IV. Multitaper Spectrum Estimate

The most used method for spectrum estimation in speech processing is the windowed periodogram known also as Hamming windowed DFT spectrum [17]. This spectrum estimate can be formulated as:

$$\hat{S}(f) = |\sum_{j=0}^{N-1} w(j)s(j)e^{\frac{-2i\pi f}{N}}|^2 \quad (1)$$

where $f \in \{0, 1, ..., N-1\}$ is the frequency bin index, $[s(0), s(1), ..., s(N-1)]$ is a speech frame with length N and $w(j)$ denotes a window function. Equation 1 is also called a single-taper periodogram.

The use of a single taper for spectrum estimate reduces the bias (the difference between the estimate spectrum and the real spectrum) but causes a problem of discarding a significant part of the signal. Indeed, the spectral estimate will have high variance. This variance can be reduced by using multitaper method for spectrum estimate. The multitaper spectrum estimate is calculated by:

$$\hat{S}_{MT}(f) = \sum_{p=1}^{M} \lambda(p)|\sum_{j=0}^{N-1} w_p(j)s(j)e^{-\frac{2i\pi if}{N}}|^2 \quad (2)$$

where w_p is the p^{th} data taper ($p = 1, 2, ..., M$), $\lambda(p)$ is the weight of the p^{th} taper and N is the frame length.
The tapers w_p are chosen to be orthonormal such as:

$$\sum_j w_p(j)w_q(j) = \delta_{pq} = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

So, the multitaper spectrum is the weighted sum of sub-spectra. The single taper power spectrum estimate can be obtained when we set $p = M = 1$ and $\lambda(p) = 1$. The windows functions (tapers) of the multitaper approaches are taken so that the sub-spectra have uncorrelated estimation errors. Indeed, a low variance estimate is obtained when we average these uncorrelated sub-spectra. In literature, different multitaper methods have been proposed: Thomson multitaper [18], Multipeak multitaper [19] and SWCE (Sinusoidal Weighted Cepstrum Estimator) multitaper [20]. The choice of tapers affect significantly the calculated spectrum estimate. These tapers should be resistant to spectral leakage.

V. Classification

Different techniques have been used in the literature to classify stressed states in speech such as Artificial Neural Networks [2], Gaussian Mixture Models [3], Hidden Markov Model [1], and Support Vector Machines [4].

Support vector machines (SVM) are widely used as learning machine classifiers in the past decade due to their best performances compared to traditional mythologies used in

solving signal processing problems. The SVM have showed to perform other well-known classifiers in stress and emotion recognition and was used in many studies [21], [22].

In this work, we are interested to SVM to classify multitaper MFCCs into appropriate stress classes. The principle of SVM is to find an optimal hyperplane that separates two classes using the maximized margin criteria. The SVM were originally implemented to solve problems of binary classification. But, real applications oblige the researchers to extend SVMs to multiclass approaches [23]. Different methods have been proposed including one-versus-rest (1vR), one-versus-one (1v1) and directed acyclic graph (DAG).

The 1vR approach is the simplest and the oldest one [24]. It builds k SVM (one per class). The 1vR SVM is based on training the j^{th} SVM with all the examples of this j^{th} class considered as positive ones and all other examples as negative ones. It can also be used to discover the reject of example which does not belong to any of the k classes. However, this approach is almost criticized because of its asymmetry due to the fact that each hyperplane is training with a number of negative examples more important than the number of positive ones. This problem can be resolved by the use of the 1v1 method which constructs $k(k-1)/2$ binary classifiers (k the number of classes) [25]. Despite that the 1v1 uses a bigger number of hyperplane in the training phase than the 1vR, this method is often faster. The DAG SVM is trained in the same way of training the 1v1 method [26]. However, a rooted binary directed acyclic graph is used during the test phase. This graph has $k(k-1)/2$ nodes where each node is a binary SVM. This method has been developed in order to resolve the problem of areas of indecision caused by the OAO approach. Moreover, it has been demonstrated that the DAG SVM is faster than the 1v1 and the 1vR methods.

The SVM approaches are based on kernels. In literature, there are some kernels which are widely used and are considered as standard kernels. These kernels are linear, polynomial and gaussian. In this work, we used a polynomial and a Gaussian kernel which are defined as follows:

$$K_{Gaussian}(x, x_i) = \exp(-\frac{\|x - x_i\|^2}{2\sigma^2}) \quad (4)$$

$$K_{Poly}(x, x_i) = (a* < x, x_i > +b)^d \quad (5)$$

VI. Experiments

A. Stress Corpus

The stressed database SUSAS (Speech under Simulated and Acted Stress) [27] consists of stressed speech samples recorded under simulated environment and acted environment. In this study, we used only the speech utterances recorded under simulated stress conditions. This domain consists of speech uttered by 9 speakers having 3 different dialects and contains 10 different stressed styles. SUSAS comprises a set of 35 aircraft communication words. Each speaker is reading these words twice. The sampling frequency is 8 Khz.

Four states of speech under stress: Neutral(N), Angry(A), Loud(Ld) and Lombard(Lb) are considered. In our experiments, speech utterances of 8 speakers are considered that to say that 2240 isolated words are used in the training and the test phases. The two thirds of data are used as training set and the third of this data is used for test.

B. Experimental setup

The performance of multitaper MFCCs for speech under stress recognition is evaluated using the SUSAS (Speech Under Simulated and Actual Stress) database. The MFCCs and MMFCCs features are extracted from the data collected from the isolated words which represent the four stress states examined.

For a comparative study, we implement the multitaper approach with different types of tapers (described in Table I) in order to extract multitaper MFCC features. These methods are Thomson, Multipeak and SWCE. The classical MFCC features are computed using the Hamming window. The multitaper methods were used by varying the number of tapers ($2 \leq p \leq 8$). In the experiments, the speech signals were segmented into frames of 10ms lengths. After that, each frame was weighted by a single taper or multitaper method. The different multitaper methods Thomson, Multipeak and SWCE were generated as described in [13].

For evaluation, we have used multiclass SVM approaches including 1vR, 1v1 and DAG. The multiclass SVMs approaches were implemented using the LIBSVM toolbox for Matlab. The different methods are evaluated with two kernels: polynomial and gaussian. The RBF kernel parameters (c, σ) are calculated using the cross-validation procedure [28] where:

$$c = [2^{-15}, 2^{-14}, ..., 2^{14}, 2^{15}] \quad (6)$$

and

$$\sigma = [2^{-15}, 2^{-14}, ..., 2^{14}, 2^{15}] \quad (7)$$

C. Results

In this study, a recognition system for stressed speech is implemented using MFCCs and MMFCCs features. Stress states are recognized with three multiclass SVM classifiers including 1vR, 1v1 and DAG approaches applied with polynomial and gaussian kernels.

The results of the first classification approach are presented in table II. Comparing the classification rates obtained with traditional MFCCs to those obtained with the multitaper MFCC features computed with the three multitaper methods, we can notice that when we use the polynomial kernel, the classification accuracies are improved ranging from 80.05% to 93.44%. These results depend on the multitaper method used and the number of tapers. However, the use of the RBF kernel with MMFCC ameliorate the performance of the stressed speech recognition system in same cases. The best classification rate of 99.44% is obtained with the application of Thomson multitaper method with a number of tapers $p = 3$.

The same experiments were conducted with the two other multiclass SVM approaches: 1v1 and DAG. Results of the second and the third SVM approaches are illustrated in table III and table IV. Indeed, for the 1v1 SVM method, the classification accuracies obtained with the polynomial kernel

TABLE I: Stress Speech Recognition Systems based on Single Taper and Multitaper MFCCs

Appraoch	Description
Hamming	single taper MFCCs using Hamming window
Thomson	Multitaper MFCCs using dpss tapering
Multipeak	Multitaper MFCCs using multipeak tapering
SWCE	MFCC are computed from sinusoidal weighted (i.e., sine tapered) spectrum estimate

TABLE II: Classification Accuracy using 1vR/SVM

Feature	Number of tapers	Polynomial	Gaussian
MFCC	1	76.94	99.06
MMFCC-Thomson	2	93.30	98.92
	3	92.40	99.44
	4	88.62	97.72
	5	89.42	99.19
	6	89.95	98.52
	7	90.46	98.61
	8	87.95	99.19
MMFCC-Multipeak	2	92.10	99.33
	3	84.47	97.32
	4	92.36	99.33
	5	80.05	98.12
	6	91.70	98.52
	7	81.84	98.96
	8	86.63	98.26
MMFCC-SWCE	2	93.44	98.79
	3	92.77	99.19
	4	91.29	98.79
	5	91.16	77.64
	6	88.89	99.19
	7	91.03	98.25
	8	87.81	98.39

are in [74.44%, 95.04%] and with the RBF kernel in [97.99%, 99.30%].

We can remark that both kernels gives important results but the improvement is very remarkable with the polynomial kernel. The accuracy has passed from 50.88% for classical MFCC to 95.05% with the Multipeak multitaper method used with a number of tapers $p = 2$. This amelioration is also obtained when we applied the DAG SVM method. The best results are obtained when we use the gaussian kernel associated to MMFCC computed with Multipeak approach ($p = 7$).

From the three tables representing the results, we can conclude that the use of multitaper methods to extract MFCC features improve the performances of the stressed speech recognition system. An important improvement exceeding 45% is achieved with the polynomial kernel but the best accuracies are often obtained by the application of the gaussian kernel with the three multiclass SVM approaches.

VII. Conclusion

In this paper, we have used the multitaper method in order to extract MFCC features for stressed speech recognition. The windowed DFT used in the traditional extraction process is replaced by multitaper spectrum estimation. The evaluation of the stressed speech recognition system is realized on SUSAS database using multiclass SVM approaches. The results show that there is an improvement in the performances of the implemented stressed speech recognition systems with multitaper MFCCs.

For future work, multitaper approach can be useful in extraction of other features from the stressed speech such as multitaper gammatone frequency cepstral coefficients and multitaper PLP in order to improve the recognition accuracy. Also, we can test the multitaper MFCC with other classification approaches.

Acknowledgment

This work has been supported by the Research Laboratory LR-SITI-ENIT (Signal, Images and Information Technolo-

TABLE III: Classification Accuracy using 1v1/SVM

Feature	Numbre of tapers	Polynomial	Gaussian
MFCC	1	50.80	98.79
MMFCC-Thomson	2	92.50	98.66
	3	91.29	99.30
	4	91.96	99.19
	5	90.62	98.39
	6	89.02	98.79
	7	88.53	99.03
	8	86.74	98.39
MMFCC-Multipeak	2	95.04	98.52
	3	75.36	97.99
	4	92.77	99.19
	5	75.10	98.92
	6	92.10	98.92
	7	74.44	99.48
	8	89.17	98.93
MMFCC-SWCE	2	92.77	99.19
	3	92.77	98.92
	4	93.03	98.92
	5	91.56	98.52
	6	89.29	99.06
	7	88.62	98.66
	8	90.36	98.92

TABLE IV: Classification Accuracy using DAG/SVM

Feature	Number of tapers	Polynomial	Gaussian
MFCC	1	49.73	99.19
MMFCC-Thomson	2	93.97	98.92
	3	93.78	99.30
	4	91.29	98.66
	5	90.36	98.39
	6	89.15	98.79
	7	87.70	99.30
	8	86.61	98.52
MMFCC-Multipeak	2	95.04	98.25
	3	78.44	99.19
	4	92.36	98.52
	5	74.96	97.99
	6	91.03	98.79
	7	75.87	99.61
	8	90.77	98.26
MMFCC-SWCE	2	93.03	98.92
	3	91.83	99.19
	4	92.23	98.52
	5	90.89	99.19
	6	91.03	99.46
	7	90.62	99.19
	8	90.49	99.06

gies), ENIT.

REFERENCES

[1] B.D. Womack , J.H.L. Hansen , N-channel hidden Markov models for combined stressed speech classification and recognition Speech and Audio Processing, IEEE Transactions on, VOL. 7, NO. 7, pp. 668-677, 1999.

[2] B. D.Womack and J. H. L. Hansen, Classification of speech under stress using target driven features, Speech Commun: Speech Under Stress, VOL. 20, pp. 131150, November 1996.

[3] D. Ververidis , C. Kotropoulos , Emotional speech classification using Gaussian mixture models, IEEE International Symposium on, VOL. 3, pp. 2871- 2874, 2005.

[4] T. Nguyen and I. Bass, Investigation of Combining SVM and Decision Tree for Emotion Classification, Proceedings of the Seventh IEEE International Symposium on Multimedia (ISM05), December 2005.

[5] P Boersma, Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics to-Noise Ratio of a Sampled Sound, Proceedings of the Institute of Phonetic Sciences, VOL. 17, pp. 97110, 1993.

[6] S. E. Bou-Ghazale and J. H. L. Hansen, A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress, IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 8, NO. 4, pp. 429-442, JULY 2000.

[7] S. Besbes and Z. Lachiri, Multi-class SVM for stressed speech recognition, 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), pp. 782-787, March 2016.

[8] C. Ittichaichareon, S. Suksri and T. Yingthawornsuk, Speech Recognition using MFCC, International Conference on Computer Graphics, Simulation and Modeling (ICGSM2012),Pattaya (Thailand), July 28-29, 2012.

[9] T. Kinnunen, R. Saeidi, J. Sandberg, M. Hansson-Sandston, What Else is New than the Hamming Window? Robust MFCCs for Speaker Recognition via Multitapering, International Speech Communication Association, INTERSPEECH, pp. 2734-2737, 2010.

[10] M. Hansson-Sandsten and J. Sandberg, Optimal cepstrum estimation using multiple windows, IEEE ICASSP, Taipei, Taiwan, pp. 30773080, 2009.

[11] M. A. Wieczorek and F. J. Simons, Minimum-variance multitaper spectral estimation on the sphere, The Journal of Fourier Analysis and Applications, vol. 13, no. 6, pp. 665692, 2007.

[12] J. Sandberg, M. Hansson-Sandsten, T. Kinnunen, R. Saeidi, P.Flandrin, and P. Borgnat, Multitaper estimation of frequency warped cepstra with application to speaker verification, IEEE Signal Processing Letters, vol. 17, no. 4, pp. 343346, 2010.

[13] T. Kinnunen, R. Saeidi, F. Sedlk, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and Haizhou Li, Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 7, SEPTEMBER 2012.

[14] Y. Attabi, M. J. Alam, P. Dumouchel, P. Kenny, and D. OShaughnessy, Multiple windowed spectral features for emotion recognition, in Proceedings of the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP13), pp. 75277531, IEEE, Vancouver, Canada, May 2013.

[15] S. V Chapaneri, D. D. Jayaswal, Multi-Taper Spectral Features for Emotion Recognition from Speech, International Conference on Industrial Instrumentation and Control (ICIC), Pune India, May 28-30, 2015.

[16] J. Trangol and A. Herrera, Traditional Method and Multi-Taper to Feature Extraction Using Mel Frequency Cepstral Coefficients, International Journal of Information and Electronics Engineering, Vol. 5, No. 1, January 2015

[17] F. J. Harris, On the use of windows for harmonic analysis with the discrete Fourier transform, Proc. IEEE, vol. 66, no. 1, pp. 5184, Jan. 1978.

[18] D. J. Thomson, Spectrum estimation and harmonic analysis, IEEE proceeding, vol. 70(9), pp. 10551096, 1982.

[19] M. Hansson and G. Salomonsson, A multiple window method for estimation of peaked spectra, IEEE Trans. on Sign. Proc., vol. 45(3), pp. 778781, 1997.

[20] K. S. Riedel and A. Sidorenko, Minimum bias multiple taper spectral estimation, IEEE Trans. Signal Process., vol. 43, no. 1, pp. 188195, Jan 1995.

[21] C.W. Hsu, C.C. Chang and C. J. Lin, A practical guide to support vector classification, Department of Computer Science and Information Engineering National Taiwan University, Taipei, Taiwan, 2009. Available: www.csie.ntu.edu.tw/ cjlin/

[22] A. Hassan and R. I. Damper, Multi-class and hierarchical SVMs for emotion recognition, In Proc. Interspeech, 2010.

[23] C. Hsu and C. Lin, A comparison of methods for multiclass support vector machines, IEEE Transactions on Neural Networks, VOL. 13, NO. 2, pp. 415-425, 2001.

[24] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and other Kernel-based Learning Methods, Cambridge, UK: Cambridge University Press, 2000.

[25] C. Hsu and C. Lin, A comparison of methods for multiclass support vector machines, IEEE Transactions on Neural Networks, Vol. 13, No. 2, pp. 415425, 2001.

[26] J. Platt, N. Cristianini, and J. Shawe-Taylor, Large margin DAGs for multiclass classification, Proceedings of Neural Information Processing Systems, NIPS99, Denver, CO, pp. 547553, 2000.

[27] J. H. L. Hansen and S. E. Bou-Ghazale, Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database, EUROSPEECH 1997.

[28] L. I. Kuncheva, Combining pattern classifiers methods and algorithms. New York: Wiley, 2004.