Whispered Speech Recognition Based on DTW algorithm and µFCC feature

Branko R. Marković and Jovan Galić

***Abstract*—This paper presents the results of normal and whispered speech recognition using the µFCC (µ-law Frequency Cepstral Coefficients) feature. This feature uses a warping frequency function and it is applied at the front-end of ASR. The Dynamic Time Warping algorithm is used at the back-end of the ASR system. All experiments were performed using the part of the Whi-Spe database. Four scenarios are analyzed: normal/normal, whisper/whisper, normal/whisper and whisper/normal in the speaker dependent mode. The results confirmed an expected improvement in recognition of whispered speech compared to the standard LFCC and MFCC features.**

Keywords— µFCC (µ-law Frequency Cepstral Coefficients); whispered speech; DTW (Dynamic Time Warping); speech recognition.

I. INTRODUCTION

THE speech has different modes and one of standard classification is: whisper, soft speech, normal (neutral), loud and shout [1]. Very interesting is whispered speech because is quietly different compared to normal, and at same time is intelligible and in most cases easy to understand. A lot of researches who were involved in normal speech recognition also are trying to apply different tools for whisper [2-4]. They made different results with more or less success.

This paper analyzes different scenarios related to whispered and normal speech. For this purpose the DTW algorithm is used with specific warping scale (µ warping**).**

The DTW (Dynamic Time Warping) algorithm [5] is known as "old" pattern matching method for back-end ASR systems. There are many different new method like HMM (Hidden Markov Models), DNN (Deep Neural Networks), SVM (Support Vector Machines) etc. but for quick and valuable compression DTW is still very successful. Many researchers use the DTW as a method for initial classification of patterns, and then use other methods for more precise results.

For this research speech patterns from the Whi-Spe database [6] are used. The database contains 10,000 patterns which are representation of 50 different words spoken in normal and whispered mode. Five male and five female volunteers were included in this recording. For recording the special acoustical room is used where noise is suppressed.

All experiments are based on three types of features: Mel-frequency cepstral coefficients (MFCC) plus delta, Linear frequency cepstral coefficients plus delta and µFCC cepstral coefficients (LFCC) plus delta. For all experiments the following training/test scenarios are examined: comparison between normal and normal patterns (N/N scenario), comparison between whisper and whisper patterns (W/W), comparison between normal and whisper patterns (N/W) and comparison between whisper and normal patterns (W/N).

The paper has the following structure: the second part explains how to obtain MFCC and LFCC feature vectors from the initial wave files. The third part explains how to obtain µFCC feature vectors. The forth part shows the results of experiments for all mentioned features. The final remarks and hints for further research are presented at the conclusion.

II. MFCC AND LFCC FEATURES EXTRACTION

Mel-frequency cepstral coefficients are traditionally very popular feature for speech characterization. The mel-frequency scale (Fig. 1) emulates human's ear perception. The frequency in mel is calculated using the following equation:

$$f[mel] = 2595 * \log_{10}(1 + f[Hz]/700) \quad (1)$$

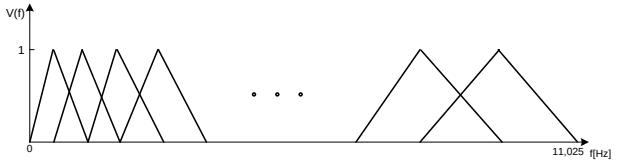

Fig. 1. Filters based on mel scale

Linear-frequency scale has the same shape for all filters (Fig. 2). The Linear frequency cepstral coefficients shows some advantage compare to MFCC in case of speaker identification in whisper [7].

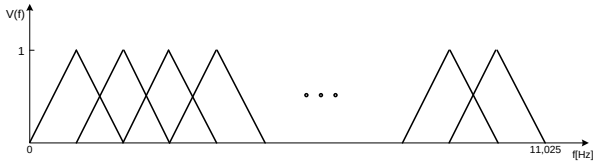

Fig. 2. Filters based on linear scale

The way to obtain MFCC and LFCC features is depicted in Fig. 3. It is a process of getting usually three types of vectors as an output: vectors of cepstral coefficients, vector of cepstral and delta cepstral coefficients and vectors of cepstral, delta

Branko R. Marković is with the Faculty of Technical Science Čačak, University of Kragujevac, Department of Computer Science and Software Engineering, Čačak, Svetog Save 65, Serbia (e-mail: brankomarko@ yahoo.com)

Jovan Galić is with the Faculty of Electrical Engineering, University of Banja Luka, Department of Electronics and Telecommunications, Banja Luka, Bosnia and Herzegovina (e-mail: jovan.galic@etf.unibl.org)

cepstral and delta-delta cepstral coefficients. For this research the first two types are used.

The inputs are wave files from Whi-Spe database [6]. All patterns are recorded with sampling rate of 22050 Hz, 16 bits per sample. Only difference between MFCC and LFCC features is in the scale which is used: MFCC counts the log energy over mel scale while LFCC counts the log energy over linear scale.

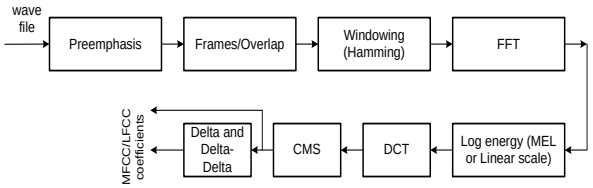

Fig. 3. Block diagram for MFCC/LFCC based features

The preprocessing assumes a several steps to get the feature vectors from an initial wave file (Fig. 3). The following steps are performed: preemphasis, framing with overlap, Hamming window, Fast Fourier Transformation (FFT), Log energy over a specific scale, DCT (Discrete Cosine Transform) and CMS (Cepstral Mean Subtraction) normalization.

The first is preemphasis block and it produces a spectrally flattened signal. Then, in the framing/overlap block, the signal as an output of preemphasis, is divided into frames. Each frame contains 512 samples, and then it is overlapped 50%. In next block the frames are weighted with the Hamming window.

The next step is the FFT, which calculates short time spectra of the signal. Then, the Log energy is calculated over the specific scale (mel or linear).

Finally, the Discrete Cosine Transformation with CMS are applied to produce the cepstral coefficients.

The CMS is a normalization method and is very important approach for whispered speech recognition [10,11].

For calculation of the first derivative (Delta), three neighboring frames are included.

Based on the mel scale and the preprocessing, two types of vectors are produced:

- vector containing 12 MFCCs and
- vector containing 24 coefficients (12 MFCCs and 12 Delta MFCCs).

Similarly, based on the linear scale the following vectors are obtained:

- vector containing 12 LFCCs and
- vector containing 24 coefficients (12 LFCCs and 12 Delta LFCCs).

These types are used in all experiments.

III. µFCC Feature Extraction

The reason to involve µFCC feature is in the following: due to unvoiced nature of whispered speech the spectrum is relatively flat. Some significant part of whispered information is in higher part of speech spectrum. The mel scale, due to its nature, is not able to "catch" these information. The linear scale shows better performances for part of higher frequencies but has worse resolution for lower frequencies. So, as a compromise, the new, warping function of frequency is involved [8,9].

This function is called µ-law and originally is involved for speech compression and expending in Japan and North America. The µ-law is defined by the following equation:

$$f_\mu = f_N \frac{\ln(1+\mu\frac{f}{f_N})}{\ln(1+\mu)} \qquad (2)$$

where $f_N = f_S/2$ (f_S is the sampling frequency), and for these experiments f_N =11025 Hz. The µ is a positive number and can have different values. For this research µ takes values {0,1,2}. Fig. 4 shows the warping functions of µ-law for these three values {0,1,2}[9]. For µ =0, the scale is linear, and practically the feature are LFCC, as mentioned before.

The µFCC feature should make some compromise between linear and mel scale with focus to improve whispered speech recognition.

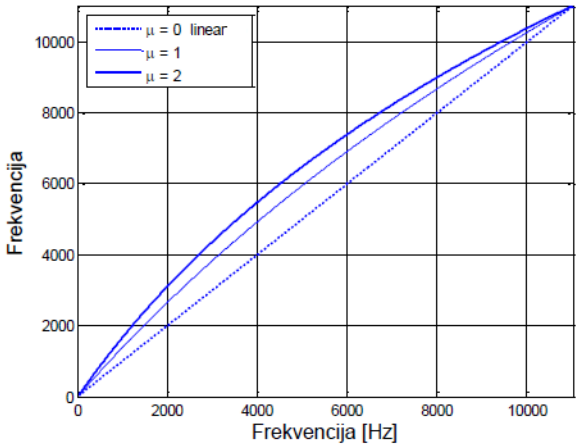

Fig. 4. Warping functions

In order to obtain µFCC features the block diagram from Fig. 5. is used. It is clone to earlier mention diagram for LFCC and MFCC feature (Fig. 3). The main difference is usage of warping function over frequencies when Log energy is calculated.

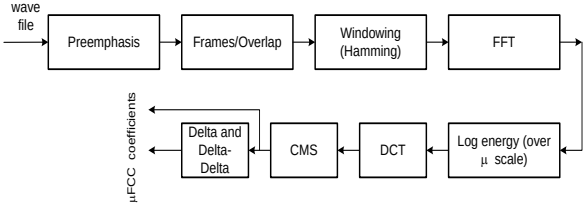

Fig. 5. Block diagram for µFCC based features

For these experiments two values for μ are used: 1 and 2. So, for μ=1 two types of vectors are considered:

- vector containing 12 μFCCs (μ-law Frequency Cepstral Coefficients) and
- vector containing 24 coefficients (12 μFCCs and 12 Delta μFCCs).

Similarly, for μ=2 two types of vectors are produced.
All these vectors are used in all experiments.

IV. Results

For the purpose of testing these different features a software package is developed using the MATLAB. There are two parts of this software: the first one converts the targeted wave files from the Whi-Spe database into the set of MFCC, LFCC, and μFCC feature vectors (two type of vectors for all of them, and for μFCC two different values of μ are used). The second part compares feature vectors using the DTW algorithm.

The DTW algorithm uses the dynamic programming and it allows finding an optimal path between the starting and ending points. The speech patterns are represented by a set of feature vectors. Two set of patterns are used: the first set of 50 patterns is used as a reference, and all other patterns (nine sets, each of 50 patterns) are used as test data. For a local constraint the type I is implemented [12]. No global constraints are used.

This research used two (of ten) speakers from Whi-Spe database: one female (Speaker1) and one male (Speaker6).

For all types of feature vectors mentioned before, the results are expressed as the Word Recognition Rate (WRR). Four scenarios in speaker dependent mode are analyzed: normal/normal (denoted as N/N), whisper/whisper (W/W), normal/whisper (N/W) and whisper/normal (W/N). Tables I and II shows results for "Speaker1" and "Speaker6" using four different vector features with 12 cepstral coefficients.

TABLE I
WORD RECOGNITION RATE FOR "SPEAKER1" USING 12 CEPSTRAL COEFFICIENTS

Scenario	LFCC	μFCC (μ=1)	μFCC (μ=2)	MFCC
N/N	98.89	99.11	99.56	99.78
W/W	95.56	96.89	97.11	97.78
N/W	81.56	84.75	85.11	78.22
W/N	67.78	66.44	64.00	46.44

TABLE II
WORD RECOGNITION RATE FOR "SPEAKER6" USING 12 CEPSTRAL COEFFICIENTS

Scenario	LFCC	μFCC (μ=1)	μFCC (μ=2)	MFCC
N/N	96.44	98.22	98.67	99.33
W/W	89.33	92.89	95.11	95.11
N/W	62.00	65.56	68.00	68.00
W/N	51.33	50.89	49.78	36.44

Tables III and IV give results for "Speaker 1" and "Speaker 6" using all mentioned feature vectors with 12 cepstral and 12 delta cepstral coefficients.

TABLE III
WORD RECOGNITION RATE FOR "SPEAKER1" USING 12 CEPSTRAL AND 12 DELTA CEPSTRAL COEFFICIENTS

Scenario	LFCC	μFCC (μ=1)	μFCC (μ=2)	MFCC
N/N	98.67	98.89	99.56	99.56
W/W	96.00	97.11	97.11	97.78
N/W	82.00	84.44	84.89	78.22
W/N	66.44	68.00	64.44	45.78

TABLE IV
WORD RECOGNITION RATE FOR "SPEAKER6" USING 12 CEPSTRAL AND 12 DELTA CEPSTRAL COEFFICIENTS

Scenario	LFCC	μFCC (μ=1)	μFCC (μ=2)	MFCC
N/N	96.22	97.78	98.44	99.11
W/W	88.67	93.11	94.22	94.89
N/W	63.11	66.44	67.33	67.56
W/N	51.33	50.89	49.33	37.56

Based on the results from Tables I and II can be concluded that the MFCC feature gives very good results for match scenarios (Normal/Normal and Whisper/Whisper). But for mismatch scenarios μFCC feature is giving better results than MFCC (about 8% for N/W scenario, 30% for W/N scenario – for Speaker1 and about 28% for W/N scenario - for Speaker6). Also, for match scenarios μFCC feature gives better result than LFCC feature for both speakers.

The results in Tables III and IV are based on vectors with 12 cepstral coefficients plus 12 delta cepstral coefficients. In some cases there are improvements related to the word recognition rate with these vectors, but they are not significant (i.e. for Speaker1 and W/N scenario, μFCC feature (μ=1) gives 1,5% better result than for cepstral).

In general, based on results from Tables I-IV it is easy to conclude that the Speaker1 has better results in all scenarios compared to Speaker6. Speaker1 is better "whisperer". Hence, on Fig. 6. the results of Speaker1 are depicted for the feature vectors which contains 12 cepstral coefficients.

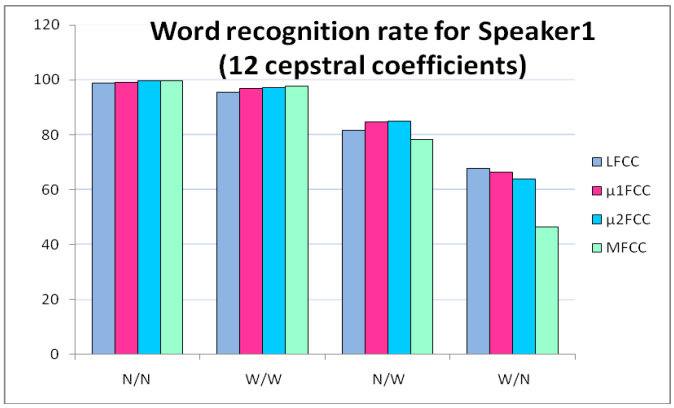

Fig. 6. WRR for Speaker1 using vectors of 12 cepstral coefficents

As it expected, the best results are for N/N, W/W, N/W and W/N scenarios, respectively. With µ1FCC the µFCC feature where µ=1, is denoted. Similarly, µ2FCC means µ=2.

It is interesting that µFCC feature for N/W scenario gives better results than LFCC and MFCC.

Fig. 7 shows results of Speaker1 for all scenarios when vectors are containing 12 cepstral plus 12 delta cepstral coefficients.

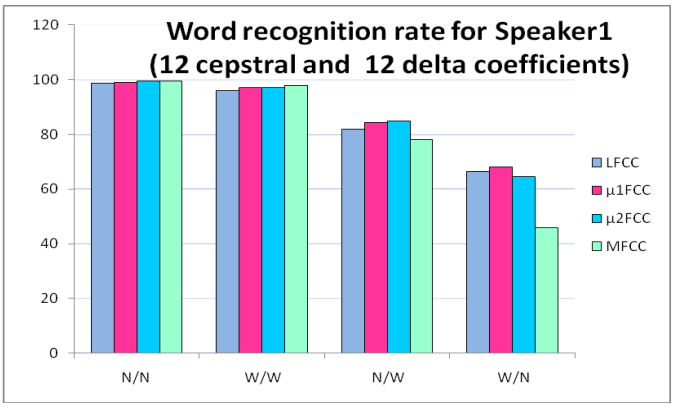

Fig. 7. WRR for Speaker1 using 12 cepstral plus 12 delta coefficents

If Fig. 6 and Fig. 7 are compared the similar trend for all scenarios and all features is evident. Only, for W/N scenario LFCC and µ1FCC changed their places.

V. Conclusion

As it expected, the best recognition results are obtained for Normal/Normal scenario and they are above 99% when MFCC feature is used. Also for Whisper/Whisper scenario the WRR is the best with MFCC. So, for match scenarios MFCC gives good results.

When mismatch scenarios are analyzed µFCC allows better results than MFCC. This is especially visible for W/N scenario where the improvement is from 28% to 30%. Obviously, these results are optimistic and give a hint to make more detailed research how the values of µ cause different WRR.

Comparing the length of feature vectors (12 cepstral coefficients vs. 24 coefficients -12 cepstral and 12 delta) the results are similar. The reason behind it can be "clean" speech in Whi-Spe database, while the delta parameters are usually efficient for noisy speech.

Further analysis may include all ten speakers from Whi-Spe database, and also more different values for µ. Instead of {0,1,2} values it can be numbers with decimal point [9]. That should provide new interesting results.

References

[1] C. Zhang, J.H.L. Hansen, "Analysis and classification of Speech Mode: Whisper through Shouted," *Interspeech 2007*, 2007, pp. 2289-2292

[2] S.T. Jovičić, Z.M. Šarić, „Acoustic analysis of consonants in whispered speech," *Journal of Voice*, 22(3), 2008, pp. 263-274.

[3] Matsuda M., Kasuya H. (1999). "Acoustic nature of the whisper", *Proc. Eurospeech 99*, 1, 1999, pp. 137-140.

[4] S.T. Jovičić, "Formant feature differences between whispered and voiced sustained vowels," *ACUSTICA - Acta Acustica*, 84(4), 1998, pp. 739-743.

[5] L. Rabiner, B-H. Juang, "Fundamentals of speech recognition", Prentice Hall, New Jersy, 1993

[6] B. Marković, S.T. Jovičić, J. Galić, Đ. Grozdić,: Whispered Speech Database: Design, Processing and Application, 16th International Conference, TSD 2013, I. Habernal and V. Matousek (Eds.): TSD 2013, LNAI 8082, Springer-Verlag Berlin Heidelberg, pp. 591-598

[7] Fan X., Hansen J.H.L. Speaker identification with whispered speech based on modified LFCC parameters and feature mapping. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009). 2009. pp. 4553–4556.

[8] Sklar B. Digital Communications: Fundamentals and Applications: 2nd edition. Prentice-Hall. 1988. 776 p.

[9] J. N. Galić, S. T. Jovičić, V. D. Delić, B. R. Marković, D. S. Šumarac Pavlović, Đ. T. Grozdić, *„HMM-based Whisper Recognition using µ-Law Frequency Warping "*, SPIIRAS Proceedings, Issue No 3(58), 2018, pp. 27-52, ISSN 2078-9181, DOI 10.15622/sp.58

[10] J. De Veth, L. Boves, "Channel Normalization Techniques for Automatic Speech Recognition over the Telephone", – Speech Communication, 25, pp. 149-164, 1998.

[11] Grozdić Đ., Jovičić S., Šumarac-Pavlović D., Galić J., Marković B. (2017). "Comparison of Cepstral Normalization Techniques in Whispered Speech Recognition", Advances in Electrical and Computer Engineering, Vol. 17. Number 1, 2017, pp 21-26.

[12] Sakoe H. and Chiba S. (1978). „Dynamic programming optimization for spoken word recognition", IEEE Trans. Acoustics, Speech, Signal Proc., pp 43-49, 1978