Recognition of Normal and Whispered Speech Based on RASTA Filtering and DTW Algorithm

Branko R. Marković, Goran Stevanović, Slobodan T. Jovičić, Miomir Mijić, Jovan Galić and Đorđe T. Grozdić

Abstract: **This paper presents the results of normal and whispered speech recognition using RASTA (Relative Spectral) filtering applied on PLP (Perceptual Linear Prediction) feature. It is used for a front-end while the DTW (Dynamic Time Warping) algorithm is used for a back-end of ASR system. All isolated words in these experiments are from the Whi-Spe database. In the experiments four training/test scenarios are analyzed: normal/normal, whispered/whispered, normal/whispered and whispered/normal. The speaker dependent mode is used with 10 speakers. The results confirm a good improvement in recognition when RASTA filtering is applied, especially in the mismatch scenarios.**

Keywords – Speech recognition; Whispered speech; Whi-Spe database; PLP; RASTA filtering; DTW algorithm.

I. INTRODUCTION

THE speech can be classified in different ways, but the basic modes are: whisper, soft, normal, loud and shouted [1]. The normal speech is a topic of researcher's investigation in many decades, but other modes, especially the whisper is still a challenge. The whisper is used in many different situations: when someone wishes to be discrete at a public place (market, library, metro, etc.), when someone wants to hide identity over the telephone calls, when someone has a health problem etc. Today, the mobile telephone service is very popular, so the whisper is a common speech mode [2]. The recognition of whispered speech is a challenge for many researches [1-4], but the research is also focused on both whisper and natural speech and their mixed scenarios.

Well known techniques for automatic speech recognition (ASR) applied to normal speech can also be used on whisper. The most popular ASRs using the following methods: DTW (Dynamic Time Warping), HMM (Hidden Markov Models) and ANN (Artificial Neural Networks) [5]. In this research the standard DTW method [6] is used as the ASR back-end. The Whi-Spe database [7] which contains 10,000 patterns of single words which are spoken in normal and whispered mode is taken for training and testing. For this database ten students (five female and five male) had read vocabulary of 50 words ten times in both modes (normal and whisper). All 10,000 patterns are recorded in a special room where the noise is suppressed.

For a purpose of this research the speaker dependent mode is considered and the entire database is used. The experiments are divided into two parts: one which applies RASTA filtering and one which doesn't. For each experiment we included four training/test scenarios: comparison between normal and normal patterns (N/N scenario), comparison between whisper and whisper patterns (W/W scenario), comparison between normal and whisper patterns (N/W scenario) and comparison between whisper and normal patterns (W/N scenario).

This paper is structured in the following way: the second part explains a PLP (Perceptual Linear Prediction) feature with RASTA filtering. This is process of extraction in order to obtain vectors from initial wave files. The third part describes the DTW method. The results in form of tables and diagram are given in the forth part. The ideas for further research and the final discussion are presented at the conclusion.

II. PLP FEATURES WITH RASTA FILTERING

The PLP speech analysis is proposed initially by Hermansky [8]. Later on, Hermansky and Morgan [9] proposed a new way "how to make speech changes less sensitive to the slowly changing of steady-state factors of speech". For that purpose the filters are implemented, and this type of filter is called RASTA (Relative Spectral). The features with RASTA can be obtained based on the block diagram from Fig. 1. This is the ASR's front-end and it requires a number of steps to get the feature vectors from an initial wave file. The feature vectors are compared in order to obtain a local distance. The wave files are from the Whi-Spe database [7]. They are recorded with sampling rate of 22,050 Hz, 16 bits pre sample. These samples (in form of wave files)

Branko R. Marković is with the School of Electrical Engineering, University of Belgrade, Blvd. Kralja Aleksandra 73; Čačak Technical College, Computing and Information Technology Department, Svetog Save 65, Čačak, Serbia (e-mail: brankomarko@ yahoo.com).

Goran Stevanović is with Railways of the Republic of Srpska a.d. Doboj, Bosnia and Herzegovina; Čačak Technical College, Computing and Information Technology Department, Svetog Save 65, Čačak, Serbia (e-mail:stevanovic.doboj@gmail.com)

Slobodan T. Jovičić is with the School of Electrical Engineering, University of Belgrade, Telecommunication Department, Blvd. Kralja Aleksandra 73; Life Activities Advancement Center, Gospodar Jovanova 35, Belgrade, Serbia (e-mail: jovicic@etf.rs).

Miomir Mijić is with the School of Electrical Engineering, University of Belgrade, Telecommunication Department, Blvd. Kralja Aleksandra 73; Belgrade, Serbia (e-mail: emijic@etf.rs).

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: jgalic@etfbl.net)

Đorđe T. Grozdić is with the School of Electrical Engineering, University of Belgrade, Telecommunication Department, Blvd. Kralja Aleksandra 73; Life Activities Advancement Center, Gospodar Jovanova 35, Belgrade, Serbia (e-mail: djordjegrozdic@gmail.com).

are inputs in the preprocessing system while the outputs are vectors of PLP cepstral coefficients (on which RASTA either applied or not).

The common steps to generate feature vectors for all types are: preemphasis, framing with overlap, windowing, Fast Fourier Transformation, Bark scale filter bank and Equal-loudness with Intensity-loudness (first two branches on Fig. 1).

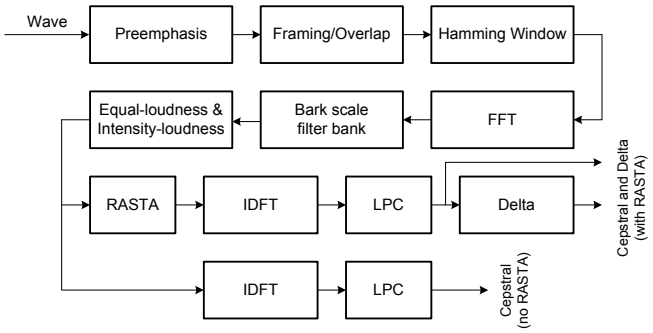

Fig. 1. Block diagram for RASTA based filtering

The goal of the preemphasis block is to produce a spectrally flattened signal and makes it less susceptible to finite precision effects later in the signal processing. In the framing/overlap block, the output signal of the preemphasis is divided into N frames, each of them with 512 samples, and they are overlapped 50%. The new frames are weighted with a Hamming window in the next block. The purpose of windowing is to taper the signal to zero at the beginning and end of each frame. The next step is the FFT (Fast Fourier Transformation), which calculates a short time spectra for the signal. After the FFT the Bark scale filter banks are applied. The frequencies in Barks are calculated based on formula:

$$f_{Bark} = 6 * \ln\left(\frac{f}{600} + \left(\left(\frac{f}{600}\right)^2 + 1\right)^{0.5}\right) \quad (1)$$

The centre frequencies in the filter bank are evenly spread on the Bark scale and they are approximately 1 Bark apart. The shapes of the filters are proposed by Hermansky [8]:

$$\psi = \begin{cases} 0 & f_{Bark} - f_{c(Bark)} < -2.5 \\ 10^{(f_{Bark} - f_{c(Bark)} + 0.5)} & -2.5 \le f_{Bark} - f_{c(Bark)} \le -0.5 \\ 1 & -0.5 < f_{Bark} - f_{c(Bark)} < 0.5 \\ 10^{-2.5(f_{Bark} - f_{c(Bark)} - 0.5)} & 0.5 \le f_{Bark} - f_{c(Bark)} \le 1.3 \\ 0 & f_{Bark} - f_{c(Bark)} > 1.3 \end{cases} \quad (2)$$

The equal-loudness is involved to approximate the sensitivity of hearing at different frequencies and it's given as:

$$E = \frac{(\omega^2 + 56.8 * 10^6)\omega^4}{(\omega^2 + 6.3 * 10^6)^2 (\omega^2 + 0.38 * 10^9)(\omega^6 + 9.58 * 10^{26})} \quad (3)$$

where ω is an angular frequency ($\omega = 2\pi f$).

The intensity-loudness compression approximates a non-linear relationship between signal intensity and perceived loudness. It is a cubic root amplitude compression given with:

$$\Phi_m = (X_m)^{0.33} \quad 1 \le m \le M \quad (4)$$

where m is filter's order and $M=30$ (number of filters).

For filtering the following an IIR filter is used:

$$H(z) = 0.1z^4 \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \quad (5)$$

This filter allows suppression of constant factors in each spectral component of the short-term auditory-like spectrum prior to go in LPC analysis.

The inverse Discrete Fourier Transformation is applied to the output of RASTA filters and then autocorrelation function is used to evaluate the Linear Prediction Coefficients (LPC) [10]. Then if Delta block is applied we obtain the delta cepstral coefficients. In order to calculate the first derivative (Delta), three neighboring frames are used.

If the RASTA not applied (the fourth branch on Fig. 1) classical PLP cepstral coefficients are produced (after IDFT and LPC blocks).

Finally, three types of vectors are produced and analyzed:

- vectors containing 12 cepstral coefficients without RASTA,
- vectors containing 12 cepstral coefficients with RASTA and
- vectors containing 24 coefficients with RASTA (12 PLPCCs and 12 Delta PLPCCs).

They are a base which we used for experiments and research.

III. DTW ALGORITHM

For a back-end of the ASR system DTW algorithm is used. In order to compare two speech patterns and to find their similarity one of the best ways is to use this algorithm [5]. It allows comparison of two patterns which are represented by the feature vectors and the beginnings and ends of these vectors should be overlapped. Fig. 2 shows one potential path which guides from the beginning point *(1,1)* to the ending point *(P, Q)* (where *P=10*, *Q=12* for this example) by usage of the certain number of steps.

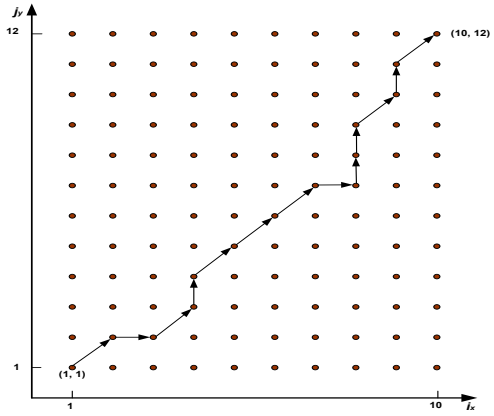

Fig. 2. Example of a DTW possible path.

We used here a local constraint of Type I proposed by Sakoe and Chiba [11]. It is depicted in Fig. 3. It shows how to reach the point *(i,j)* form the previous states. Weight from *(i-1, j-1)* to *(i, j)* is denoted by *k* and can be changed.

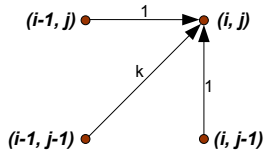

Fig. 3. Local constrain of Type I.

The DTW algorithm gives ability to find an optimal path using dynamic asynchronous programming. The key idea is to use the recursion steps and to find all local paths that reach a state *(i,j)* in exactly one step from the three potential previous states (*(i-1,j)* or *(i-1,j-1)* or *(i, j-1)*). The appropriate algorithm can be implemented through the next steps [5]:

1) Initialization:

$$D(1,1) = d(1,1) * m(1) \tag{6}$$

2) Recursion:

$$D(i, j) = \min_{i', j'} \left[D(i', j') + d((i', j'), (i, j)) \right] \tag{7}$$

3) Termination:

$$d(P, Q) = \frac{D(P, Q)}{M_{\Phi}} \tag{8}$$

where *d(i,j)* is a local distance between vectors *i* and *j*, *D(i,j)* is an accumulated distance for the global path up to the point *(i,j)* and M_{Φ} is a normalization factor.

IV. Results

In order to test speech recognitions a software package is developed using the MATLAB. The software converts all wave files from the Whi-Spe database into the sets of PLP-RASTA feature vectors (three type of vectors mentioned above). Then these vectors are compared using DTW algorithm in speaker dependent mode.

The speech patterns (words of colors, word of numbers and acoustically balanced words [12]) are represented by a set of vectors. The first set of patterns (50 words) is used as a reference, and the other patterns (nine sets, each consisting of 50 words) as test data. For a local constraints the type I proposed by Sakoe and Chiba [11] is used where diagonal step is preferred. Global constraints are not used.

For all three types of vectors, the results expressed as the word recognition rate (WRR) for normal/normal (N/N), whisper/whisper (W/W), normal/whisper (N/W) and whisper/normal (W/N) scenarios are provided in Tables I-IV. for each particular speaker. The first five speakers (denoted with “Speaker 1”,..., ”Speaker 5”) are female, and the last five (“Speaker 6”,..., ”Speaker 10”) are male.

TABLE I
WORD RECOGNITION RATE FOR N/N SCENARIO (%)

Feature /Speaker	PLP cepst. without RASTA	PLP cepst. with RASTA	Cepst. + Δ (with RASTA)
Speaker 1	99.78	99.33	99.33
Speaker 2	99.33	100	100
Speaker 3	98.44	97.33	97.33
Speaker 4	99.33	99.33	99.56
Speaker 5	99.11	99.78	99.78
Speaker 6	98.00	98.22	98.89
Speaker 7	96.22	98.89	98.89
Speaker 8	98.67	98.67	98.89
Speaker 9	98.89	99.11	99.33
Speaker 10	92.67	99.11	98.89
Avrg	**98.04**	**98.98**	**99.09**

TABLE II
WORD RECOGNITION RATE FOR W/W SCENARIO (%)

Feature /Speaker	PLP cepst. without RASTA	PLP cepst. with RASTA	Cepst. + Δ (with RASTA)
Speaker 1	94.89	98.44	98.44
Speaker 2	96.89	98.44	98.67
Speaker 3	97.56	98.89	98.89
Speaker 4	94.67	98.22	98.67
Speaker 5	93.33	96.67	96.67
Speaker 6	77.11	87.78	88.22
Speaker 7	90.89	95.33	95.78
Speaker 8	86.48	93.11	93.56
Speaker 9	93.78	98.44	98.44
Speaker 10	81.33	90.22	89.78
Avrg	**90.69**	**95.55**	**95.71**

TABLE III
WORD RECOGNITION RATE FOR N/W SCENARIO (%)

Feature /Speaker	PLP cepst. without RASTA	PLP cepst. with RASTA	Cepst. + Δ (with RASTA)
Speaker 1	56.00	76.89	77.56
Speaker 2	31.56	39.56	38.67
Speaker 3	49.78	84.44	85.78
Speaker 4	44.89	68.00	68.67
Speaker 5	40.44	57.56	55.78
Speaker 6	31.78	65.33	66.00
Speaker 7	47.78	70.89	70.44
Speaker 8	56.44	70.22	71.78
Speaker 9	64.00	77.78	81.56
Speaker 10	35.33	67.78	69.56
Avrg	**45.80**	**67.84**	**68.58**

TABLE IV
WORD RECOGNITION RATE FOR W/N SCENARIO (%)

Feature /Speaker	PLP cepst. without RASTA	PLP cepst. with RASTA	Cepst. + Δ (with RASTA)
Speaker 1	54.44	70.44	69.78
Speaker 2	32.22	34.44	34.22
Speaker 3	43.78	64.89	64.67
Speaker 4	44.44	45.33	46.00
Speaker 5	42.00	45.56	46.44
Speaker 6	33.11	47.56	46.67
Speaker 7	44.00	54.89	54.89
Speaker 8	29.33	61.11	62.00
Speaker 9	48.44	66.44	68.00
Speaker 10	26.67	56.00	56.44
Avrg	**39.84**	**54.67**	**54.91**

Based on Table I (for normal/normal scenario) we can see the best recognition is with 24 cepstral coefficients when RASTA is applied and the WRR is more than 99%. The result is worse for about 1% when RASTA is not applied.

For whisper/whisper scenario recognition is between 95% and 96% when RASTA is applied, but it is about 90.70% (worse 5% - 6%) when RASTA is not applied (Table II).

For mismatch scenarios (results in Table III and Table IV) when RASTA is applied the recognition is 68.5% (for normal/whisper scenario) and 56% for whisper/normal scenario. By applying RASTA the recognition is improved up to 23% (for normal/whisper scenario what is respectable).

Fig. 4 shows results for these four scenarios where RASTA is applied and where is not.

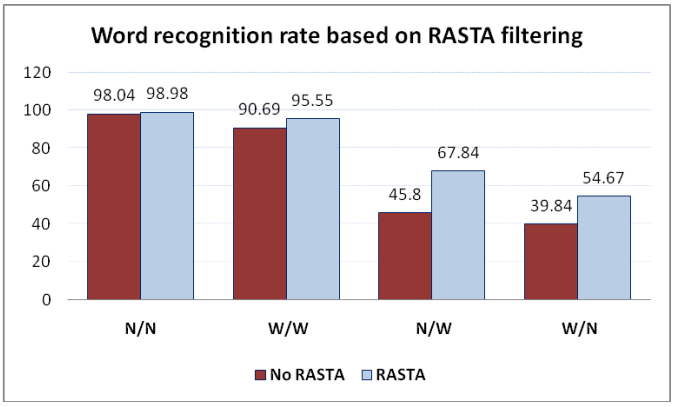

Fig. 4. Results of words recognition based on RASTA filtering.

V. CONCLUSION

The best recognition results are obtained for match scenarios. For normal/normal scenario with cepstral coefficients and RASTA filtering the WRR is around 99% and it was expected. Then, the result for whisper/whisper scenario for same type of vectors is also high – around 96%. In both cases RASTA filtering made an improvement.

For mismatch scenarios better results are for normal/whisper scenario than for whisper/normal scenario. Usage of RASTA improves word recognition rate more then 15% for W/N scenario and more then 23% for N/W scenario.

The results when 12 cepstral coefficients are used (with RASTA) vs. when 24 coefficients are used (12 cepstral and 12 delta) are similar, but is a little bit better when delta is used. The reason for this can be usage of "a clean speech", because delta and delta-delta parameters are mostly efficient when a speech has noise in it.

Further analysis may include different local and global constraints for DTW algorithm and also these results can be compared to PLP features with CMS normalization [13]. These may give new light and interesting hints for more challenges and research.

REFERENCES

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

[2] J T. Ito, K. Takeda, F. Itakura, "Analysis and Recognition of Whispered speech," *Speech Communication*, 2005, pp. 129-152.

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

[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, J. Galić, Đ. Grozdić, S. T. Jovičić, "Application of DTW method for whispered speech recognition", Speech and Language 2013, 4th International Conference on Fundamental and Applied Aspects of Speech and Language, Belgrade, October 25-26, 2013.

[7] 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

[8] H. Hermansky, "Perceptual linear predictive (PLP) analysis for speech", *J. Acoust. Soc. Am.*, pp.1738-1752, 1990.

[9] H. Hermansky and N. Morgan, "RASTA processing of speech", IEEE Trans. on Speech and Audio Proc., vol. 2, no. 4, pp. 578-589, Oct. 1994.

[10] B. R. Marković and Đ. T. Grozdić, "The LPCC-DTW Analysis for Whisperd Speech Recognition", Proceedings of 1st International Conference of Electrical, Electronic and Computer Engineering, IcETRAN 2014, pp. AKI1.1.1-4, Vrnjačka Banja, Serbia, June 2-5, 2014.

[11] H. Sakoe, S. Chiba, "Dynamic programming optimization for spoken word recognition", IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-26(1):43-49, February 1978.

[12] S. T. Jovičić, Z. Kašić, M. Đorđević, M. Rajković, "Serbian emotional speech database: design, processing and evaluation", *SPECOM-2004*, St. Petersburg, Russia, 2004, pp. 77-81.

[13] B. R. Marković, S. T. Jovičić, M. Mijić, J. Galić and Đ. T. Grozdić, "Recognition of Whispered Speech Based on PLP Features and DTW Algorithm", Proceedings of 3rd International Conference on Electrical, Electronic and Computing Engineering, IcETRAN 2016, pp. AK 1.3.1-4, Zlatibor, Serbia, June 13-16, 2016.