Baltic J. Modern Computing, Vol. 3 (2015), No. 1, 29-42

Modified Filterbank Analysis Features for Speech Recognition

Deividas ERINGIS, Gintautas TAMULEVIČIUS

Vilnius University, Institute of Mathematics and Informatics,
Akademijos St. 4, LT-08863, Vilnius, Lithuania

deividas.eringis@mii.vu.lt, gintautas.tamulevicius@mii.vu.lt

Abstract: Auditory model based feature systems include filterbank analysis and nonlinear compression of the speech signals. The Mel Frequency Cepstral Coefficients (MFCC) is the state-of-the-art feature system employing this auditory model. In this paper we proposed to modify MFCC analysis by applying power nonlinearity operator instead of logarithmic and to modify the size of filterbank. Power nonlinear operator caused increased recognition rate of deteriorated speech by 2.4 %. In combination with reduced filterbank size (down to 20 filters) power nonlinearity enhanced robustness of speech recognition: the gain of recognition rate varied from 0.6 % to 3 % in comparison with common MFCC features for different noise levels.

Keywords: speech recognition; Mel frequency cepstral coefficients; band-pass filters; nonlinearity coefficient; power nonlinearity.

1. Introduction

Various techniques have been proposed for noise robust and speaker independent speech recognition. Most of the state-of-the-art recognition systems exploit knowledge of acoustical signal processing in human auditory system for feature extraction. In 1980 Davis and Mermelstein introduced Mel Frequency Cepstral Coefficients (MFCC), the combination of good discrimination capability with low computational complexity made this method one of the most widely used features extraction method for speech recognition (Huang et al., 2001; Dong et al., 2008; Kinnunen et al., 2012). Nevertheless it has been shown that MFCC is highly affected by noises and the lack of robustness in adverse recordings, because the features are easily corrupted by distortions or noise (Kim et al., 1999; Ishizuka et al., 2006; Dimitriadis et al., 2011).

Various MFCC analysis enhancements were introduced for more robust and efficient speech analysis: increased filterbank size (Yang et al. 1992; Ganchev et al. 2005), the use of different shape filterbanks (gammatone, reversed triangular) (Kim and Stern, 2012; Chakroborty et al., 2006). Regardless of various proposals to improve MFCC it seems that the biggest obstacle for robust operation are additive noises, adhere speech, channel imbalance etc. In this paper we explore different filterbank modifications to improve speech recognition.

2. Auditory model based speech signal processing

2.1. Auditory model

It is considered that the deeper knowledge and understanding we have on human auditory system, the better we will get at constructing noise robust and speaker independent feature system for the description of the linguistic content of the speech signal. Various psychoacoustic investigations were performed in order to determine how the speech signal is processed in human auditory system (Fletcher, 1938, 1940; Stevens and Volkman, 1940; Zwicker et al., 1957). Profound inner human ear analysis (Allen, 1985; Yang et al. 1992, Ghitza, 1994) has shown that basilar membrane (a part of inner ear) can be simulated using a bank of filters with the following nonlinear operator after it (see Fig. 1). The nonlinear compression of signals is performed to convert filterbank output to pattern for further transmission to so called neural converters – inner hair cells (IHC). The further process of conversion and processing of the speech signal is still unknown thus modelling of the human auditory system is restricted to implementation of above mentioned filterbank and nonlinear operator.

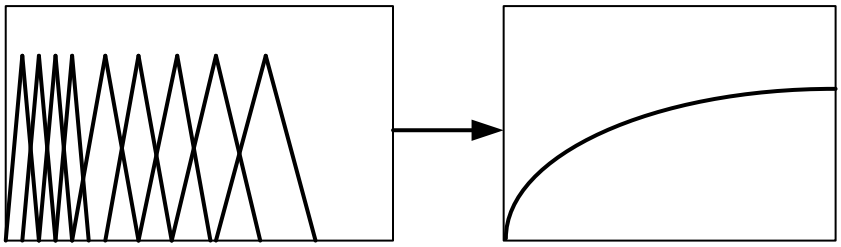

Fig. 1 Computational model in auditory system.

The filterbank is implemented as the bank of overlapped bandpass filters with the defined arrangement in frequency scale. The uniform arrangement of filters in logarithmic scale is the most common technique, which is supposed to simulate perception in human auditory system.

Another arrangement approach is based on critical bands. The term “critical band” was first introduced in research on loudness and relation to human hearing process (Fletcher, 1938). The energy of the signal is quasi-linearly integrated with respect to loudness and masking ability within critical band (Greenberg and Ainsworth, 2004). Several options for scale of critical band have been applied: Mel scale and Bark scale. These two methods are used commonly for developing non-uniform filterbanks for speech recognition purposes (Rabiner and Juang, 1993, Shannon and Paliwal, 2003, Dimitriadis et al., 2011).

In the following chapters we will overview the properties of power nonlinearity and characteristics of filterbank as separate units of auditory model.

2.2. Implementation of the auditory model

Davis and Mermelstein (1980) have proposed a new feature system for speech recognition – The Mel-Frequency Cepstrum Coefficients (MFCC). Features are derived

from real cepstrum of a short-time windowed speech signal. Signal is approximated in a nonlinear frequency scale – Mel scale (Stevens and Volkman, 1940). This scale is shown to have similar approximation capabilities as human auditory system.

In the classic model of MFCC the filterbank is implemented as a set of triangular band-pass filters arranged in non-uniform frequency scale. It has been found that perception of a given frequency F_0 in human auditory is effected by energy in a critical band around the given frequency F_0 (Allen, 1985). In lower frequency range up until 1 kHz central frequencies of critical bands are modelled as increasing linearly and then increasing logarithmically. The Figure 2 gives an example of triangular bandpass filterbank. Generally the filterbank size (*FBs*) varies from 20 to 40 filters, without any clear physical/fundamental reasoning.

There were proposed various filterbank modifications to enhance speech analysis process. Applying various parameter values of filterbank analysis for speaker verification (Ganchev et al. 2005) has shown that the increase of filterbank size provides slightly better speaker identification. Some researchers reason that various shapes of filter function (overlapped triangular – default setting for MFCC, side-by-side triangular, trapezoid) (Zheng et al. 2001) in MFCC is not significant for clean speech recognition, nevertheless there are no results of noisy speech recognition. Other researchers decoupled filter bandwidth from other filterbank design parameters, gaining ability to adjust bandwidth of filters (Skowronski and Harris, 2002, 2003). Experiment of different frequency warping techniques (Shannon and Paliwal, 2003) has shown that both Bark and Mel scales performed almost equally and outperformed the uniformly spaced filterbank.

In MFCC nonlinear processing is implemented using logarithm operation, nevertheless MFCC features does not exhibit threshold behaviour. Therefore, mathematical operation of logarithm for low power speech segments can produce significant changes in output with minor changes in the input (Tyagi and Wellekens 2005; Kim and Stern, 2009). Because of environmental noises, this quality becomes significant when speech signal is deteriorated, and even very small fluctuations of additive noise can produce considerable differences in the output of nonlinearity element.

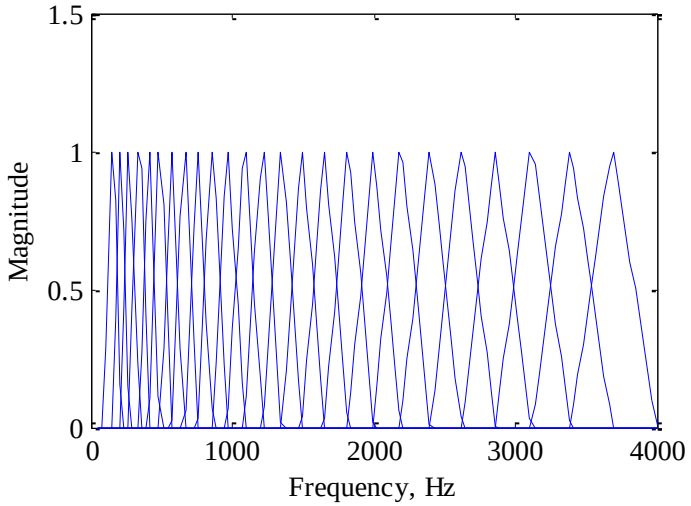

Fig. 2 Triangular shaped critical band filterbank used in calculation of MFCC

Power function was proposed for nonlinear operation also. Power nonlinearity based signal analysis approaches as perceptual linear prediction (PLP) (Hermansky, 1990) relative spectra perceptual linear prediction (RASTA-PLP) (Hermansky et al., 1992), power normalized cepstral coefficients – PNCC (Kim and Stern, 2009, 2012) were introduced for speech recognition.

The main idea of perceptual linear prediction was to implement human auditory like processing of speech signal. PLP includes following steps: calculation of short time speech power spectrum, warping of power spectrum to Bark-scale (scale models a response of human auditory system). The scaled spectrum is convolved with critical-band curve by Fletcher (1940).

The samples of critical-band power spectrum are pre-emphasized by simulated equal-loudness curve, afterwards power spectrum is processed by loudness compression – mathematical operation of cubic root. The last steps are the approximation of the power spectrum by an all-pole model (Makhoul, 1975) and cepstral analysis. Some researcher's (Hermansky, 1990; Sárosi et al., 2011) stated that the use of cubic root is more effective for noisy speech recognition than the logarithmic compression.

RASTA-PLP includes additional processing of slow spectral changes of degraded speech signal. Authors have replaced critical-band spectrum with spectral estimate, where each frequency component is filtered with infinite impulse response bandpass filter. Using this operation slow varying steady-state elements are suppressed thus, producing spectral estimate, which is less susceptive to slow spectral changes.

Analysis of power-normalized cepstral coefficients (PNCC) (Kim and Stern, 2009, 2012) also applies power law based nonlinearity element:

$$y = x^{a_0}, \tag{1}$$

where a_0=0.1. This value was experimentally proven to be approximate fit to the physiological rate-intensity function (Kim and Stern, 2009) and it was experimentally proven to outweigh logarithmic nonlinearity in MFCC processing.

We can observe similarities within these analysis approaches. Usually logarithm operation or power function is applied for nonlinear processing. Analytical formulation and evaluation of the particular auditory model based feature extraction technique is very complicated task. Thus evaluation of the feature set is performed on experimental basis commonly.

3. Work purpose and experimental setup

In this chapter we will analyse two relevant issues. We will investigate how the size of band-pass filterbank and the nonlinear operator affect speech recognition process.

3.1. Work purpose

To summarize all previous discussed feature extraction methods one can state, that all these methods basically have two processing stages: analysis in band-pass filterbank and nonlinear processing of the speech signal (see Fig. 1).

Filterbank and power nonlinearity modifications for speech recognition were thoroughly investigated. We have selected MFCC features for the experimental analysis as the state-of-the-art feature system. MFCC analysis uses 24 band-pass triangular filters and logarithm operation for auditory modelling. The filters are aligned linearly in Mel

frequency domain. The number of band-pass filters in auditory models according to various authors is much bigger and goes up to 190 (Yang et al. 1992; Ghitza, 1994). Higher number of filters can give better evaluation of the signal spectral properties in time thus giving higher robustness to noise. Thus one of the aims of our study is the effect of filterbank size on speech recognition rate. Filterbank with *M* filters (*m = 1, 2 ,..., M)*, where *m* – triangular filter is implemented using following formula:

$$H_m[k] = \begin{cases} 0 & , \ k < f[m-1], \\ \frac{2(k-f[m-1])}{(f[m+1]-f[m-1])(f[m]-f[m-1])}, & f[m-1] \le k \le f[m], \\ \frac{2(f[m+1]-k)}{(f[m+1]-f[m-1])(f[m+1]-f[m])}, & f[m] \le k \le f[m+1], \\ 0 & , \ k > f[m+1]. \end{cases} \quad (2)$$

where $H_m[k]$ – critical band filter, k – sample of a signal, $f[m]$ – boundary points uniformly spaced in Mel-scale:

$$f[m] = \frac{N}{F_S} B^{-1}\left(B(f_L) + m\frac{B(f_H)-B(f_L)}{M+1}\right), \quad (3)$$

where F_S – sampling frequency in Hz, N – length of the analysed signal in samples, B – Mel-scale frequency and B^{-1} is an inverse of a Mel-scale frequency:

$$B(f) = 1127\ln(1 + f/700), \quad (4)$$

where f – linear frequency in Hz.

$$B^{-1}(b) = 700\left(e^{b/1127} - 1\right), \quad (5)$$

where b – frequency in Mels.

After passing short-time speech signal through a filterbank we obtain a response with sharpened spectra. Then the logarithmic compression is performed, i.e. the dynamic range of a spectrum is minimized. In original paper (Davis and Mermelstein, 1980) this is done by computing log-energy of the each filter output:

$$S[m] = ln\left[\sum_{k=1}^{N} |X_a[k]|^2 \cdot H_m|k|\right], \qquad 1 \le m \le M, \quad (6)$$

where $X_a[k]$ is a power spectrum of short time speech signal $x[n]$.

Afterwards Mel frequency cepstrum coefficients are calculated using cosine transform:

$$C_{MFCC}[n] = \sum_{m=1}^{M} S[m]cos\left(\pi n\left(\frac{m+1/2}{M}\right)\right), \qquad 1 \le n \le M, \quad (7)$$

where M – number of filters. Usually the first 13 cepstrum coefficients are used with mathematically derived 26 coefficients (delta and acceleration coefficients).

In this study we decided to use power function of the form A^x. Because of human nature to perceive intensity of physical stimulus via power law (Stevens, 1957). It has been shown that human sensations (loudness, brightness, taste etc.) has a power law dependence, which is based on magnitude of physical stimulation and its perceived strength. We believe that power function can provide higher and variable degree of compression than logarithm function. The cepstral part of the MFCC analysis will be performed as follows:

$$C_{MFCC}[n] = DCT[\{|DFT(x[n])|\}^{\propto}], \quad (8)$$

where $C_{MFCC}[n]$ – cepstral feature sequence, DCT and DFT – discrete cosine and Fourier transform operations respectively, $x[n]$ – speech signal sequence, α – variable power value.

PLP and RASTA-PLP technique applies power of 0.33 claiming it as simulating the power law of hearing. PNCC technique applies experimentally obtained value of 0.1. The difference is fairly large and hardly explainable. We will modify MFCC features by varying number of band-pass filters and power value in search of noise robust and speaker independent feature set.

3.2. Experimental setup and results

There is no clear statement for physical or fundamental reasoning for the particular number of filters in critical-band structure. Preliminary assumption of the maximum number of filters may be bounded to sampling frequency of the speech.

The number of filters is set by experimental knowledge and not defined by some physical constant, this is also true for power-law element. Therefore, in this paper we will show our experimental results on modified filterbank analysis parameters, namely the number of overlapped triangular filters and experimental seek for optimal value of power-law nonlinearity.

For our experimental study we have selected the task of recognition of 111 Lithuanian words:

- training dataset was set of ~1 hour of speech recordings of 19 males and 19 females (4218 utterances totally);
- testing dataset consisted of 2 subsets: clear speech recorded by 3 males and 3 females (666 utterances) and deteriorated speech (low amplitude, high signal-to-noise ratio, yet noise-free speech) of another 3 males and 3 females (666 utterances);
- Hidden Markov models based recognizer with word level models was employed for experiment. Every word was modelled using different number of states depending on word's grapheme number (1 grapheme – 3 states);
- the length of analysis windows was set to 20 ms, window shift – 12 ms;
- recordings were captured using 8,000 Hz sampling frequency;
- overall filterbank bandwidth was set 50 – 4,000 Hz. Filters were spaced linearly in Mel frequency domain;
- filterbank size varied from 5 to 95 with the step of 5;
- non-linearity power coefficient (α) varied from 0.01 to 0.1 with the step of 0.01 (filterbank size was 26 in this case).

Our first investigation was to check whether the increment of filterbank size had a considerable impact on speech recognition. Similar experiments with various levels of additive white noise (30 dB, 20 dB, 10 dB) were executed. Results are presented in Figure 3.1 and Figure 3.2., Results of MFCC analysis are given for reference purposes.

We can see that the increase of filterbank size from 5 to 15 filters had a positive effect on recognition rate (SR) of deteriorated and clear speech. Recognition for deteriorated speech in the case of 10 dB noise level was increased by 8.56 %, for 20 dB SR was increased by 19.67 %, 30 dB – by 16.07 %. Recognition rate of noise-free deteriorated speech was increased by 9.01 %. Recognition rate of clear speech with added 10 dB noise level was increased by 29.73 %, speech with 20 dB noise level had an increase of 21.23 %, 30 dB noise level – 3.6 %, clear speech without added noise increased by 1.8 %.

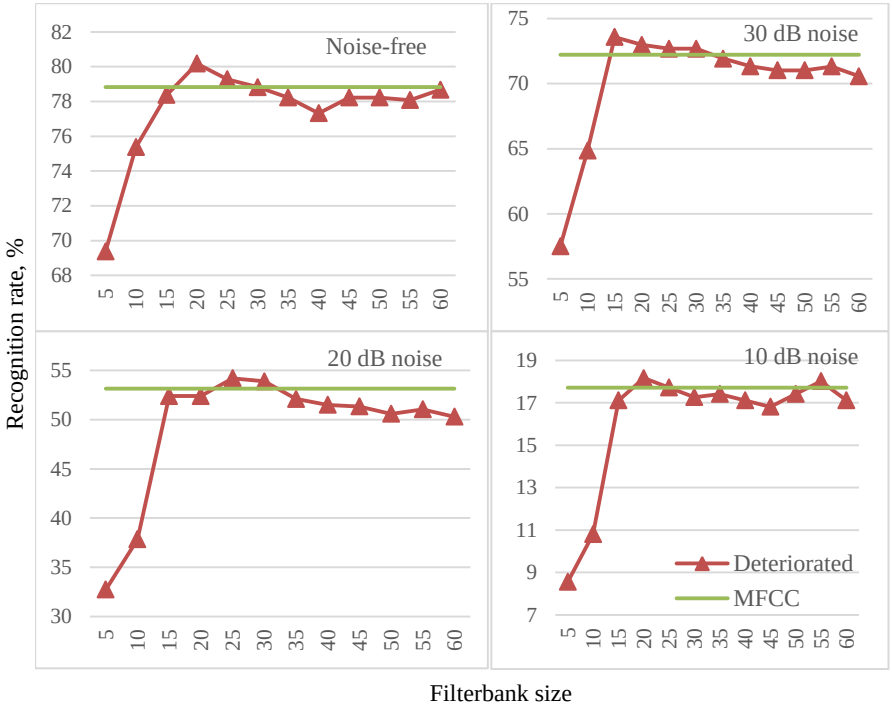

Fig. 3.1 Recognition rate of deteriorated speech using different filterbank size

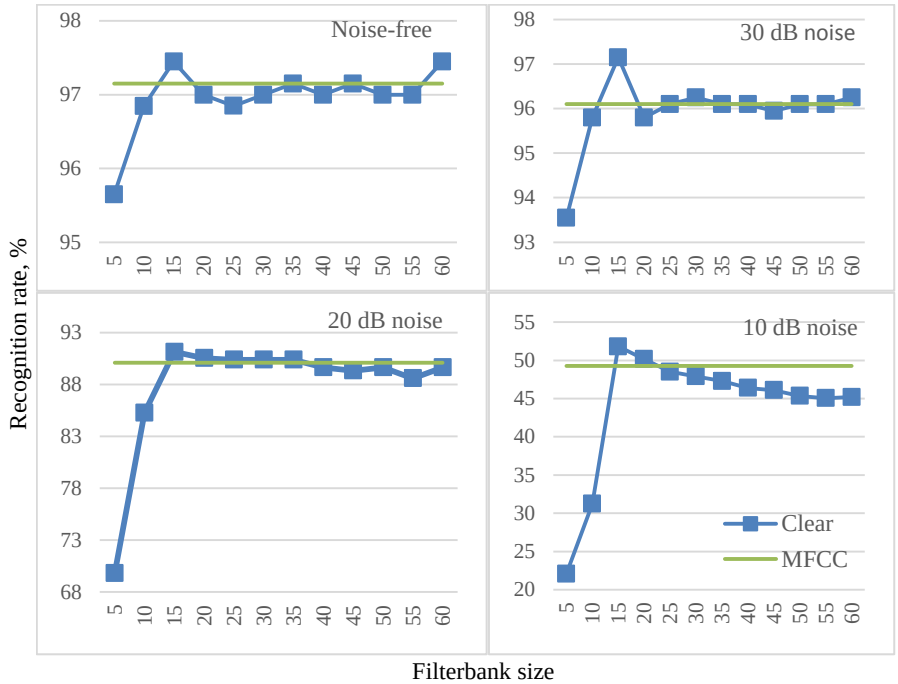

Fig. 3.2 Recognition rate of clear speech using different filterbank size.

In comparison with common MFCC features SR for deteriorated speech with added 10 dB noise level increased by 0.45 % (20 filters); for 20 dB noise SR increased by 1.05 % (25 filters); for 30 dB noise – 1.35 % (15 filters); for noise-free case a 1.35 % increase was recorded (20 filters). For clear speech the highest recognition improvement was achieved using 15 filters: with 10 dB noise – 2.55 %, with 20 dB and 30 dB noise levels – 1.05 % for noise-free speech – 0.30 %.

The further increment of filterbank size up to 95 filters did not give any SR improvement. In contrary, SR fluctuated, even decreased (in case of clear speech with 10 dB SNR noise). The results of this experiment suggest that over increasing number of filters may not improve the rate of speech recognition.

The aim of the second experiment was to investigate the effect of power-law nonlinearity on the recognition rate. Firstly we have decided to determine effective range of power-index values α. We have selected initial range of [-1; +1] for pilot study of clear speech recognition. Results of this experiment are given in Figure 4.

We can see a mirror effect in Figure 4, i.e. results are similar both for positive and negative values of power index α. Even though results are alike, we must stress that positive values of α had higher impact on speech recognition rate. Therefore, we have used value range α = [0.01; 0.1] for our next experiment.

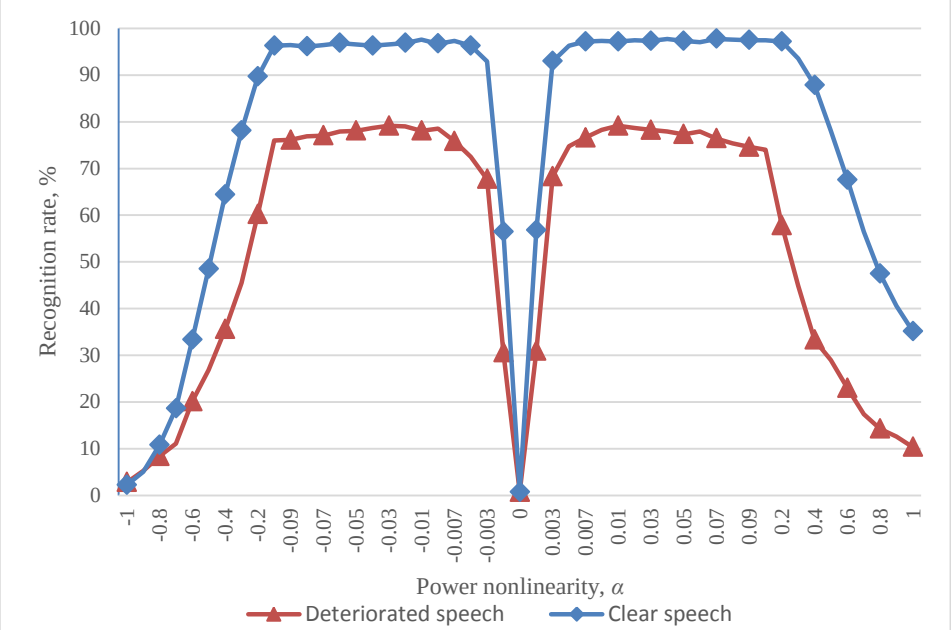

Fig. 4 Speech recognition rate dependence on power nonlinearity element α.

During this experiment we tried to determine the power index value a giving the highest recognition rate of clear and deteriorated speech with various levels of noises. Results are presented in the Figure 5.1 (for deteriorated speech) and Figure 5.2 (for clear speech).

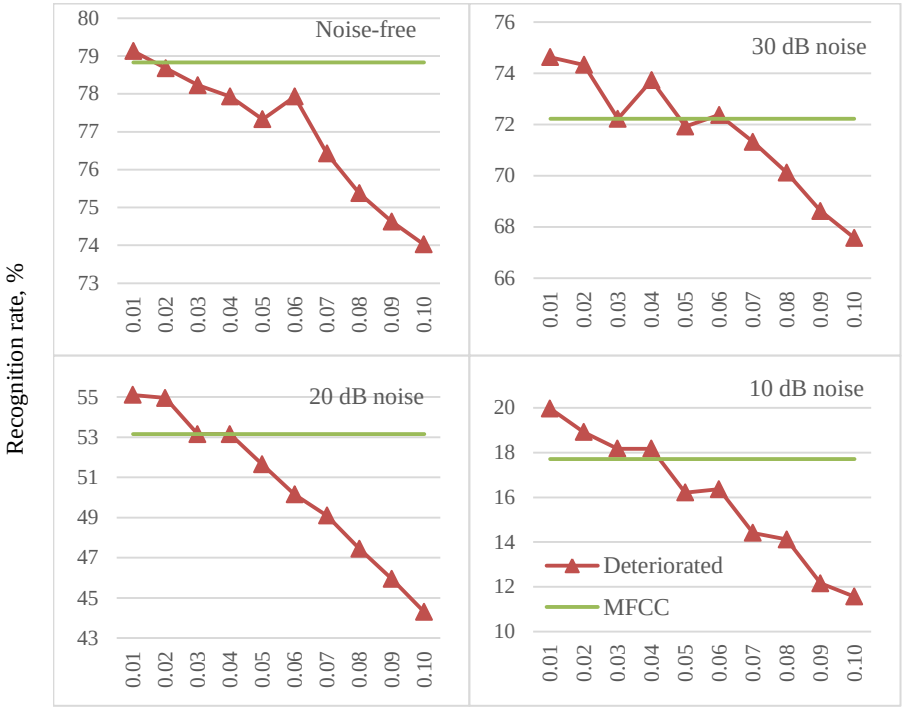

Fig. 5.1 Power nonlinearity influence on recognition of deteriorated speech.

Fig. 5.2 Power nonlinearity influence on recognition of clear speech.

The results in Figure 5 show the same tendency as in the Figure 4 (positive side of α value). The increment of α reduces recognition rate of deteriorate speech. The case of clear speech indicates fluctuation of speech recognition but it is approximately the same in regard to α variation. For deteriorated speech signal SR drops as the value of power nonlinearity element increases. The higher noise level we have the faster drop-off of recognition rate of deteriorate speech we can observe. In the clear speech case SR rate fluctuates as α increases, there is no noticeable drop-off in recognition rate. In generally, the highest speech recognition rate is achieved using α value of 0.01. In comparison with MFCC, the use of α increased deteriorated speech recognition rate from 0.30 % (added 30 dB noise level) to 2.25 % (10 dB noise level), the use of α for speech with no deteriorations increased SR by 0.9 % (10 dB and 30 dB noise levels respectively), and recognition with added 20dB noise level was increased by 0.6 %.

Selection of particular α and filter bank size (*FBs*) values produced higher speech recognition rate than common MFCC features. However, we ought to examine how these two parameters influence SR when they are combined together. For this experiment we have selected and combined previously examined values of power-law nonlinearity (a = 0.01) and the range of filterbank size ([15; 30]), which were proven to be reliable in speech recognition. Figure 6 contains obtained results for clear and deteriorated speech signals with various levels of noise.

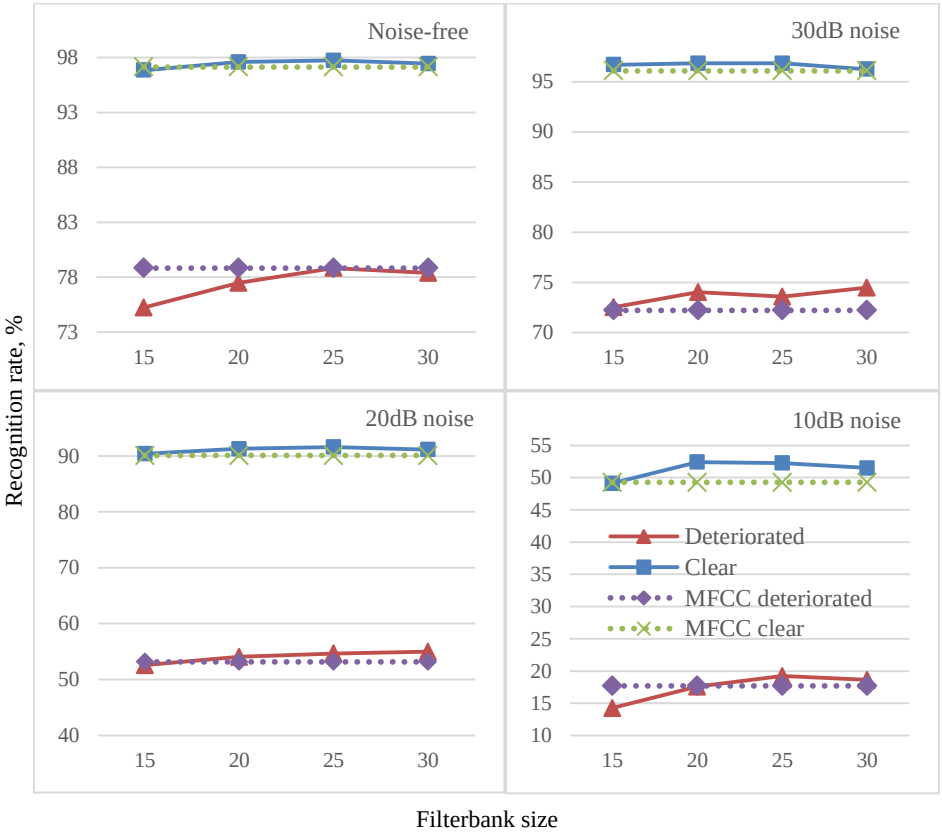

Fig. 6 Power nonlinearity with varying filterbank size effect on recognition rate of clear and deteriorated speech

A presumption that a combination of both filterbank modifications will increase speech recognition is reasonable because we can spot that the increase of filterbank size (from 15 to 25) improved recognition of deteriorated speech with a noise level of 10 dB (in other cases of noise levels recognition improved also). Recognition of clear speech shows similar response to combination of selected values of α and *FBs*. For the case when α=0.01 the optimal filterbank size is 20-25.

For comparison and generalization purposes we put all our results into Table 1. Darker cells indicate the best obtained result for particular noise level for a given speech quality. A combination of different values of α and *FBs* is presented as 0.01/20 and 0.01/25 respectively.

Table 1. Results of different techniques

Type	Deteriorated speech				Clear speech			
	10 dB	20 dB	30 dB	Noise-Free	10 dB	20 dB	30 dB	Noise-Free
MFCC	17,72	53,15	72,22	78,83	49,25	90,09	96,10	97,15
α=0.01	19,97	55,11	74,63	79,13	50,15	90,69	97,00	97,15
FBs=20	18,17	52,40	72,98	80,18	50,15	90,54	95,80	97,00
FBs=25	17,72	54,20	72,67	79,28	48,50	90,39	96,10	96,85
α/FBs 0.01/20	17,57	54,05	74,03	77,48	52,40	91,29	96,85	97,60
α/FBs 0.01/25	19,22	54,66	73,58	78,83	52,25	91,59	96,85	97,75

Recognition of clear speech using both modifications simultaneously (*FBs* and α) appear of similar capacity as using classical MFCC features. The use of power nonlinearity excels MFCC in case of deteriorated speech recognition: recognition of speech with 10 dB noise level was increased by 2.5 %, with 20 dB noise level – by 1.96 %, in case of 30 dB level – by 2.4 %. However, recognition of deteriorated speech (with no additive noise) using power based nonlinearity had a slight SR increase of 0.3 % and was indistinguishable from common MFCC feature case. Thus, application of power nonlinearity in signal analysis can improve robustness of the speech recognition process.

The use of modified size filterbank did not improve deteriorated speech recognition significantly, for example, in the case of 20 dB noise SR was lower by 0.75 %, but rate of noise-free speech recognition was increased by 1.35 %. The differences in recognition rates using different filterbank sizes were very small and thus, negligible.

Combination of α and *FBs* modifications had uneven effect for deteriorated speech. Modification "0.01/25" was superior to the MFCC analysis in deteriorated speech recognition. The improvement ranged from 0 % (noise-free speech) to 1.5 % (for 10 dB and 20 dB SNR cases). Combination "0.01/20" gave inconsistent results: recognition rate varied from -1.35 % (for noise-free speech) up to 1.8 % (for speech with 30 dB noise level).

An employment of power nonlinearity for clear speech recognition had minor and negligible improvements: from 0.6 % to 0.9 %. The use of different *FBs* had different results: for 20 filters the noisier the speech got, the better recognition rate improvement was. A 30 dB case resulted in a 0.3 % decrease, for 20 dB case an increase of 0.45 % was observed, similarly for a 10 dB case a 0.9 % increase was recorded.

The biggest improvement of SR was obtained using a modification "0.01/20" (α/*FBs*) – 3.15 % increase for a signal with 10 dB noise level. The lower noise level got the smaller speech recognition improvement was observed: 1.2 % for 20 dB

noise level speech, 0.75 % for 30 dB level and 0.45 % for noise-free speech. Similar results can be observed with the modification "0.01/25": for 10 dB noisy speech SR was increased by 3.0 %, for 20 dB the increase of 1.50 % was observed, 0.75 % for 30 dB and 0.6 % for noise-free speech.

In summary we can state that the usage of power nonlinearity based feature extraction outperforms standard Mel-frequency cepstral features in recognition of deteriorated speech by 2.4 %. The combination of power nonlinearity and the modification of filterbank size also outperforms MFCC. However selection of optimal *FBs* should be considered. Employment of power nonlinearity for recognition of clear speech did not improve SR significantly, neither did variation of *FBs*. However, the combination of both modifications caused recognition rate increase over 3 % for very noisy speech and did not make any significant change in recognition rate for noise-free speech. Thus, applying proposed MFCC modification for speech recognition one can expect improved recognition of noisy speech and unaffected recognition of noise-free speech.

4. Discussion and Conclusions

In this paper, the influence of filterbank size (*FBs*) and power nonlinearity (α) for speech recognition has been investigated. Two cases of speech quality were analyzed: deteriorated and clear signals.

Firstly we have established the value of power nonlinearity index α=0.01 giving the highest recognition rate. Variation of power index had no particular influence for clear speech recognition: power nonlinearity performed similarly as logarithmic nonlinearity. In case of deteriorated speech signals power nonlinearity outperformed logarithmic nonlinearity: recognition rate of speech with 10 dB noise level was improved by 2.40 % in comparison with common MFCC features.

A combination of different *FBs* and α gave inconsistent results. Despite this, we observed that the noisier the speech signal was analysed, the larger recognition rate improvement we got using proposed modifications. The most significant improvement was for speech with a noise level of 10 dB SNR, recognition was improved approximately by 3 % in comparison with common MFCC analysis. Modification [α=0.01; FBs=25] improved recognition of deteriorated speech by 1.5 %.

References

Allen, J. (1985). Cochlear modeling. *ASSP Magazine, IEEE* , vol.2, no.1, 3-29.

Chakroborty, S., Roy, A., and Saha, G. (2006). Fusion of a Complementary Feature Set with MFCC for Improved Closed Set Text-Independent Speaker Identification. *Industrial Technology, 2006. ICIT 2006. IEEE International Conference on* , 387-390.

Davis, B. S., Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 28, No. 4, 357-366.

Dimitriadis, D., Maragos, P., Potamianos, A. (2011). On the Effects of Filterbank Design and Energy Computation on Robust Speech Recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, vol.19, no.6, 1504-1516.

Dong, Y., Li D., Droppo, J., Jian, W., Gong, Y. and Acero, A. (2008). Robust Speech Recognition Using a Cepstral Minimum-Mean-Square-Error-Motivated Noise Suppressor. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, 1061-1070.

Fletcher, H.(1938). Loudness, masking, and their relation to the hearing process and the problem of noise measurement, *The Journal of the Acoustical Society of America*, vol. 9, 275-293.

Fletcher, H. (1940). Auditory Patterns. *Review of Modern Physics*, vol. 12, 47-65.

Ganchev, T., Fakotakis, N., Kokkinakis, G., (2005). Comparative evaluation of various MFCC implementations on the speaker verification task. *International Conference on Speach and Computer (SPECOM'05)*, vol. 1, 191–194.

Ghitza, O. (1994). Auditory models and human performance in tasks related to speech coding and speech recognition. *Speech and Audio Processing, IEEE Transactions on*, vol.2(1), 115-132.

Greenberg, S. and Ainsworth, W. (2004).*Speech Processing in the Auditory System: An overview*. In: (Speech Processing in the Auditory System 2004), 28-29

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of Acoustic Society of America*, vol. 87, no. 4, 1738-1752.

Hermansky, H. and Morgan, N. (1994). RASTA Processing of Speech. *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, 578-589.

Huang, X., Acero, A., and Hon, H. W. (2001). *Spoken Language Processing: a guide to theory, algorithms, and development*. Prentice-Hall, New Jersey.

Ishizuka, K., Nakatani, T., and Minami, Y., and Miyazaki, N. (2006). Speech feature extraction method using subband-based periodicity and nonperiodicity decomposition. *The Journal of the Acoustical Society of America*, vol. 120, no. 1., 443-453.

Kim, C. and Stern, R. M. (2009). Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction. *INTERSPEECH*, 28–31.

Kim, C., and Stern, R. M. (2012). Power-normalized cepstral coefficients (PNCC) for robust speech recognition. *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 4101-4104.

Kim, D. S., Lee, S. Y., and Kil, R. (1999). Auditory processing of speech signals for robust speech recognition in realworld noisy environments. Speech and Audio Processing, IEEE Transactions on, vol. 7, 55–69.

Kinnunen, T., Saeidi, R., Sedlak, F., Kong, Aik L., Sandberg, J., Hansson-Sandsten, M. and Haizhou, Li. (2012). Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, vol., 20, no. 7, 1990-2001.

Makhoul, J.(1975). Linear prediction: a tutorial review. *Proceedings of the IEEE*, Vol.63, 561-580.

Rabiner, L. R. and Juang, B.H. (1993). *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs: New Jersey.

Sárosi, G., Mozsáry, M., Mihajlik, P., and Fegyó, T. (2011). Comparison of feature extraction methods for speech recognition in noise-free and in traffic noise environment. *6th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 1-8.

Shannon, B. J. and Paliwal, K. K. (2003). A comparative Study of Filter Bank Spacing for Speech Recognition. *Microelectronic Engineering Research Conference*.

Skowronski, M. D. and Harris, J. G. (2002). Increased MFCC filter Bandwidth for Noise-robust Phoneme Recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 801-804.

Skowronski, M. D. and Harris, J. G. (2003). Improving the Filter Bank of Classic Speech Feature Extraction Algorithm. *IEEE Intl Symposium on Circuits and Systems*, vol. 4, 281-284.

Stevens, S. S. (1957). On the psychophysical law. *The Psychological Review*, vol. 64, 153-181.

Stevens, S. S., and Volkman, J. (1940). The relation of pitch to frequency. *American Journal of Psychology*, vol. 53, 329-353.

Tyagi, V. and Wellekens, C. (2005). On Desensitizing the Mel-cepstrum to Spurious Spectral Components for Robust Speech Recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. (ICASSP '05)*, vol.1, 529-532.

Yang, X., Wang, K., and Shamma, S.A. (1992). Auditory representations of acoustic signals, Information Theory, *IEEE Transactions on* , vol.38, no.2, 824-839.

Zheng, F., Zhang, G., and Song, Z. (2001) Comparison of Different Implementations of MFCC. *J. Computer Science & Technology*, vol. 16, no. 6. 582-589.

Zwicker, E., Flottorp, G. and Stevens, S.S. (1957). Critical Band Width in Loudness Summation, *The Journal fo the Acoustical Society of America*, vol 29, no. 5, 548-558.

Received February 27, 2015, accepted March 4, 2015