Journal of Computer Sciences

Original Research Paper

Feature Extraction Method for Improving Speech Recognition in Noisy Environments

Youssef Zouhir and Kaïs Ouni

Research Unit: Signals and Mechatronic Systems, SMS, UR13ES49,
National Engineering School of Carthage, ENICarthage, University of Carthage, Tunisia

Article history
Received: 20-12-2015
Revised: 30-01-2016
Accepted: 24-03-2016

Corresponding Author:
Youssef Zouhir
Research Unit: Signals and Mechatronic Systems, SMS, UR13ES49, ENICarthage, University of Carthage, Tunisia
Email: youssef.elc@gmail.com

Abstract: The paper presents a feature extraction method, named as Normalized Gammachirp Cepstral Coefficients (NGCC) that incorporates the properties of the peripheral auditory system to improve robustness in noisy speech recognition. The proposed method is based on a second order low-pass filter and normalized gammachirp filterbank to emulate the mechanisms performed in the outer/middle ear and cochlea. The speech recognition performance of this method is conducted on the speech signals in real-world noisy environments. Experimental results demonstrate that method outperformed the classical feature extraction methods in terms of speech recognition rate. The used Hidden Markov Models based speech recognition system is employed on the HTK 3.4.1 platform (Hidden Markov Model Toolkit).

Keywords: Feature Extraction, Peripheral Auditory Model, Hidden Markov Models, Noisy Speech Recognition

Introduction

The Automatic Speech Recognition (ASR) system, at its most elementary level, encompasses different methods drawn from research in a wide variety of disciplines and areas such as signal processing, statistical pattern recognition, linguistics and communication theory. Each of these developed methods converts the speech signal waveform to some type of parametric representation which contains relevant information capable of distinguishing between different speech sounds (Rabiner and Juang, 1993).

The conventional feature extraction methods are based on classical signal processing techniques as the linear prediction or the filter banks (Perdigao and Sá, 1998). These methods such as Mel-Cepstre (or Mel frequency cepstral coefficients) (Davis and Mermelstein, 1980) and Perceptual Linear Prediction (PLP) (Hermansky, 1990) are most used for speech recognition systems does not perform well in noisy environments, while the human auditory system is able to recognize speech in the presence of noise (Haton *et al*., 2006).

A great deal of research has been interested in a noise-robust feature extraction, particularly the Gabor feature (Missaoui and Lachiri, 2014) or the auditory features based on a new auditory model in order to improve the performance of automatic speech recognition and the feature robustness in noisy conditions (Shao *et al*., 2009).

Patterson *et al*. (1987) have modeled the frequency analysis accomplished by the human cochlea as gammatone filterbank which is popular used in Computational Auditory Scene Analysis (CASA) systems filtering (Wang and Brown, 2006). In addition, the gammachirp filter was proposed by Irino and Patterson (1997) as an extension of the gammatone filter. It was designed to generate an asymmetric gammatone-like filter by modulating the carrier-tone term of the gammatone analytic impulse response in frequency (Meddis *et al*., 2010). This characteristic of gammachirp filter was inspired by the fact the basilar membrane impulse response is frequency modulated (Irino and Patterson, 1997; 2006; Unoki *et al*., 2006).

Many developed features was incorporated the gammachirp filter in order to improve robustness of ASR under additive noise. Among them, the PLPGc feature which integrated the Gammachirp in conventional PLP framework, proposed in (Zouhir and Ouni, 2013). The RCGCC feature developed by Alam *et al*. (2014) was obtained by incorporating a bank of compressive gammachirp (Patterson *et al*., 2003) and applying sigmoid power term for mapping.

Science Publications

The PLPaGc feature proposed in (Zouhir and Ouni, 2014), includes the use of a Gammachirp Filterbank (GcFB) and outer and middle ear filtering.

A feature extraction method named as Normalized Gammachirp Cepstral Coefficients (NGCC) for noise robust speech recognition is presented in this study. The proposed method incorporates an auditory periphery model to improve recognition performance in noisy environments. Specifically, it includes the use of a second-order low-pass filter which modeled the human outer/middle ear sound transmission (Van Immerseel and Martens, 1992) and a normalized gammachirp filterbank that represents human cochlear modeling. The used filterbank consisting of 34 normalized gammachirp filters (Zouhir and Ouni, 2014; 2015), where the filters' centre frequencies are equally spaced in ERB-rate scale (Glasberg and Moore, 1990; Moore, 2012) from 50 to 8000 Hz. The HTK speech recognizer based on the Hidden Markov Models (HMM) is used for the recognition task (Young *et al.*, 2009). Each speech word is modeled by five states whole-word HMM models with a four component Gaussian mixtures for state emitting probability density. The experimental results of speech recognition in real-world noisy environments demonstrate that the proposed NGCC feature extractor provides better results compared to the Mel-Cepstre and PLP.

This paper is structured as follows: The description of the classic feature extraction method (Mel-Cepstre) is briefly presented in section 2. Section 3 details the peripheral auditory model that simulates the mechanisms performed in auditory filter system. Section 4 proposes a new feature extraction method based on this peripheral auditory model. Section 5 presents the experimental results in noisy environments. Finally, Section 6 renders some conclusions.

The Standard Mel-Cepstre

The Mel-Cepstre (or Mel frequency cepstral coefficients) method is the most widely used for speech recognition systems.

The block diagram of the major processing steps describing the computation the coefficients of Mel-Cepstre, is illustrated in Fig. 1.

This method begins with a simple short-time spectral analysis which consists to calculate the short-term amplitude spectrum for each windowed segment using Discrete Fourier Transform (DFT). It is passed through a Mel-scale filterbank consisting of triangular band-pass filters equally spaced in Mel frequency scale. The Mel-filterbank outputs are log compressed and sent to a Discrete Cosine Transform (DCT) for decorrelating the resulting coefficients. The outputs DCT coefficients designated to as Mel-Cepstre coefficients (Davis and Mermelstein, 1980).

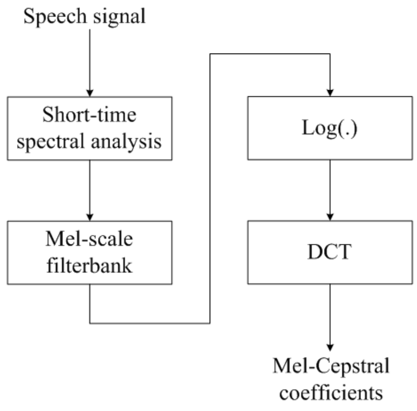

Fig. 1. Block diagram of Mel-Cepstre method

The Peripheral Auditory System Model

The peripheral auditory model is the mathematical model used to simulate the auditory mechanisms and signal processing performed in the cochlea and the outer and middle ear. The function of the outer and middle ear is to perform the filtering of the captured sound waves in order to increase the pressure of this sound (Van Immerseel and Martens, 1992). This filtering is simulated using a second-order low-pass filter by means of the transfer function defined in Equation 1, a bilinear transformation and the selection of a resonance frequency of 4 kHz (Martens and Van Immerseel, 1990; Van Immerseel and Martens, 1992):

$$H(s) = \frac{\omega_r^2}{s^2 + 0.33\omega_r s + \omega_r^2} \tag{1}$$

where, $f_r = 2\pi/\omega_r$ is the resonance frequency.

The cochlea spectral behavior is simulated by a bank of gammachirp auditory filters. The latter has been largely used to obtain a good approximation of both psychophysical and physiological data pertaining to the basilar membrane frequency selectivity in a cochlea (Irino and Patterson, 2006; Patterson *et al.*, 2003; Meddis *et al.*, 2010).

The gammachirp filterbank has an impulse response as (Irino and Patterson, 1997):

$$g_c(t) = at^{n-1}e^{-2\pi bERB(f_c)t}e^{j2\pi f_c t + jc\ln(t) + j\varphi} \tag{2}$$

where, time $t > 0$,f_c,φ,a and c are the asymptotic frequency, the initial phase, the amplitude and the chirp rate (or a parameter for the frequency modulation) respectively. 'ln' represents the natural logarithm, n and b are the two parameters that define the gamma distribution envelope and the ERB(f_c) which represents the equivalent rectangular bandwidth of the gammachirp filter at f_c, describes the critical bandwidth of human psychoacoustics. The ERB expression is defined by the following equation ERB(f_c) = 24.7+0.108 f_c in Hz.

The gammachirp filter response defined in the frequency domain is given by (Irino and Patterson, 1997; 2006):

$$\left|G_c(f)\right| = \frac{a\left|\Gamma(n+jc)\right|e^{c\theta}}{(2\pi)^n\left[(bERB(f_c))^2+\left(f-f_c\right)^2\right]^{\frac{n}{2}}} \qquad (3)$$

where, $\Gamma(n+jc)$ represents the complex gamma distribution and $\theta = \text{arctg}\left(\frac{f-f_c}{bERB(f_c)}\right)$.

The centre frequencies of gammachirp filters are distributed according to the Equivalent Rectangular Bandwidth (ERB) rate scale. The latter is an approximately logarithmic and relates to the ERBs number, ERBrate(f), such as (Glasberg and Moore, 1990; Moore, 2012; Wang and Brown, 2006):

$$ERBrate\left(f\right) = 21.4\log_{10}\left(\frac{4.37f}{1000}+1\right) \qquad (4)$$

The Proposed Feature Extraction Method

The computational process of the proposed Normalized Gammachirp Cepstral Coefficients (NGCC) is analogous to the Mel-Cepstre extraction method (Fig. 2).

The speech signal is first windowed into short frames, using a Hamming window of 25 ms with a frame shift of 10 ms. This signal can be assumed to be stationary over short intervals, thus facilitating the spectro-temporal analysis of signal and increasing the efficiency of the process of the feature extraction. The DFT is then applied for each short frame to obtain the short-term power spectrum. Subsequently, the result is processed by applying a second order low-pass filter and a normalized gammachirp auditory filterbank consisting of the frequency responses of the 34 gammachirp filters (Zouhir and Ouni, 2014). The centre frequencies of the filter are equally spaced on the ERB-rate scale from 50 to 8 kHz (sampling frequency = 16 kHz) (Glasberg and Moore, 1990; Moore, 2012). The low-pass filter is used to simulate the outer/middle ear filtering, while the gammachirp filterbank aims at simulating the cochlea spectral behavior. The logarithmic-compressed filterbank outputs are then obtained by applying the logarithmic function 'Log' in order to model loudness perception in the human auditory (Davis and Mermelstein, 1980). Finally, the Discrete Cosine Transform (DCT) is used to decorrelate the obtained outputs, yielding the Normalized Gammachirp Cepstral Coefficients (NGCC).

$$NGCC_m = \sqrt{\frac{2}{N}}\sum_{k=1}^{N} Log(X_k)\cos\left[\frac{\pi m}{N}\left(k - \frac{1}{2}\right)\right] \qquad (5)$$

$$m = 1, 2, 3, ..., M$$

where, N is the number of auditory filterbank channels, M is the number of NGCC coefficients and Log(X_k) represents the logarithmic energy output of the kth filter (k = 1, 2,...., N). N and M are chosen as the following: N = 34 and M = 12 for the NGCC computations.

Experiments

This section presents the experimental results conducted to compare the performances of the proposed feature extraction method with those of classical techniques on an Automatic Speech Recognition (ASR) task in the presence of the ambient background noises.

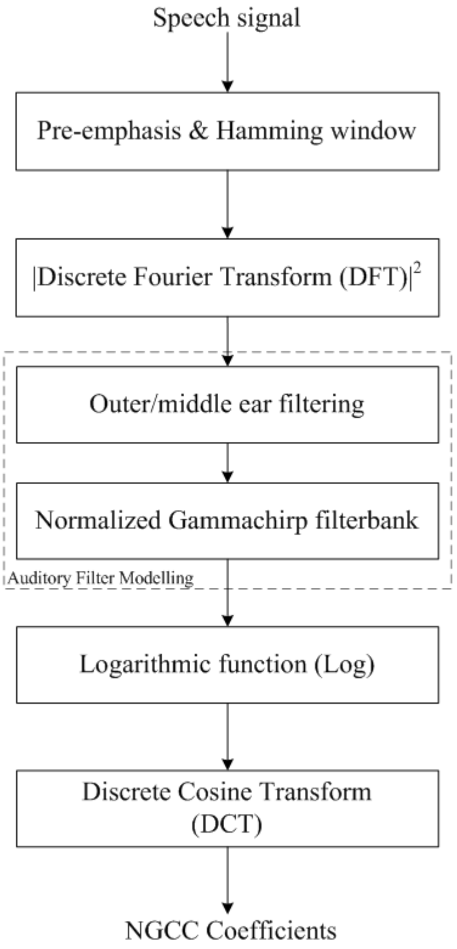
Speech signal
Pre-emphasis & Hamming window
|Discrete Fourier Transform (DFT)|^2
Outer/middle ear filtering
Normalized Gammachirp filterbank
Auditory Filter Modelling
Logarithmic function (Log)
Discrete Cosine Transform (DCT)
NGCC Coefficients

Fig. 2. The structure of the NGCC feature extraction method

Experimental Setup

The Hidden Markov model toolkit (HTK 3.4.1) (Young *et al*., 2009) is employed for building HMM based isolated word recognizers. One hidden Markov model with Gaussian Mixture density (HMM-GM), five states (a simple left-to-right model) and four diagonal mixtures per state was trained for each isolated word.

The proposed NGCC method was tested on a computer with 2.30 GHz Intel Core i3 processor, 4Go RAM and Windows 7.

The default parameter values of the gammachirp auditory filter are used in all experimental and are defined as follows: $a = 1$; $b = 1.019$; $c = 2$; $n = 4$ and $\varphi = 0$.

The used Databases

The TIMIT database (Garofolo *et al*., 1990) is taken as basis for evaluates the robustness of our method on an ASR task. It is composed of speech signals down sampled to 16000 Hz of 630 speakers (females and males) from eight major dialect regions of the United States. Each one of the speakers saying ten sentences.

In the experimental study, we used 13227 isolated-words manually extracted from the TIMIT database; 9702 isolated-words used in the training phase, while 3525 isolated-words used for the evaluation phase of ASR system. The noisy speech used in the testing phase was created by adding to the extracted isolated-words different noise types (Babble noise, Restaurant noise, Train station noise and Air-port noise) at different Signal-to-Noise Ratios (SNRs) values ranging from 0 to 15 dB. The used noises were taken from the AURORA database (Hirsch and Pearce, 2000). Figure 3 shows the temporal representations and their spectrograms of all used noises.

Results and Discussion

In all of our experiments, the speech signals samples are windowed with a Hamming analysis window into 25 ms long frames with an overlap of 10 ms. For each frame, a static feature vector consisted of 12 coefficients is computed. This vector is combined with energy (*E*), along with differential coefficients; the 1st order (Δ) and the 2nd order (*A*), to yield a feature vector of 39 coefficients for each feature extraction method (NGCC, Mel-Cepstre and PLP).

Table 1 to 4 summarize the recognition rate results obtained using the proposed NGCC feature and baselines feature (Mel-Cepstre and PLP) for the four noises (Babble noise, Restaurant noise, Train station noise and Air-port noise) at four SNR values (0, 5, 10 and 15 dB).

The results reported in these tables, showed that the proposed NGCC feature is more robust than the Mel-Cepstre and PLP feature in all noise conditions. The NGCC feature gives the better recognition results at all SNR levels, particularly for low SNR values. In the case of Babble-noise (Crowd of people) at 0 dB SNR, as an example, the recognition rate of the NGCC is respectively higher than that of the MFCC and PLP by 12.88 and 11.20.

Table 1. Comparison of recognition rates of the NGCC, Mel-Cepstre, PLP feature with babble-noise (Crowd of people) at various SNR's using HMM with 4 Gaussian Mixture densities

	SNR Levels			
	0 dB	5 dB	10 dB	15 dB
NGCC feature	49.02	76.20	90.13	95.32
Mel-Cepstre feature	36.14	64.71	84.77	92.03
PLP feature	37.82	66.33	84.20	91.40

Table 2. Comparison of recognition rates of the NGCC, Mel-Cepstre, PLP feature with restaurant-noise at various SNR's using HMM with 4 Gaussian Mixture densities

	SNR Levels			
	0 dB	5 dB	10 dB	15 dB
NGCC feature	42.70	75.89	90.64	95.12
Mel-Cepstre feature	34.18	64.82	84.99	93.82
PLP feature	35.60	65.25	85.05	92.82

Table 3. Comparison of recognition rates of the NGCC, Mel-Cepstre, PLP feature with train station-noise at various SNR's using HMM with 4 Gaussian Mixture densities

	SNR Levels			
	0 dB	5 dB	10 dB	15 dB
NGCC feature	60.48	84.43	94.52	96.54
Mel-Cepstre feature	47.74	76.11	90.84	95.32
PLP feature	47.18	77.13	89.76	94.67

Table 4. Comparison of recognition rates of the NGCC, Mel-Cepstre, PLP feature with air-port-noise at various SNR's using HMM with 4 Gaussian Mixture densities

	SNR Levels			
	0 dB	5 dB	10 dB	15 dB
NGCC feature	43.94	71.94	90.35	95.46
Mel-Cepstre feature	36.96	64.71	86.92	93.99
PLP feature	37.33	64.99	87.04	93.73

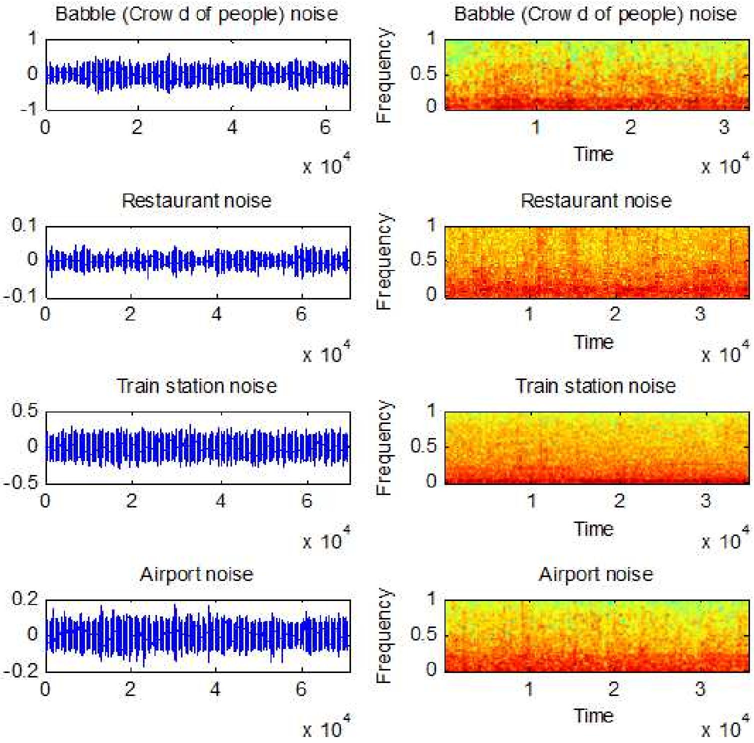

Fig. 3. The temporal representation and their spectrograms of all used noises

Conclusion

A noise robust feature extraction method was presented in this study. The proposed method is based on an auditory filter model which includes both a second order low-pass filter and a normalized auditory gammachirp filterbank to simulate the mechanisms performed in the outer/middle ear and cochlea. The bandwidth and the centers frequencies of gammachirp filterbank were determined by the critical band Equivalent Rectangular Bandwidth (ERB) and ERB-rate scale expressions respectively. Our method was tested on speech signals corrupted by real-world noises in terms of speech recognition rate. It was shown that the proposed method outperforms the other conventional method like Mel-Cepstre and PLP.

Author's Contributions

Both authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Alam, M.J., P. Kenny, P. Dumouchel and D. O'Shaughnessy, 2014. Robust feature extractors for continuous speech recognition. Proceedings of the 22nd European Signal Processing Conference, Sept. 1-5, IEEE Xplore Press, Lisbon, pp: 944-948.

Davis, S.B. and P. Mermelstein, 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process., 28: 357-366. DOI: 10.1109/TASSP.1980.1163420

Garofolo, J., L. Lamel, W. Fisher, J. Fiscus and D. Pallett *et al*., 1990. DARPA, TIMIT acoustic-phonetic continuous speech Corpus. National Institute of Standards and Technology, Technical Report No. NISTIR 4930, Gaithersburg, MD, Speech Data Publish on CD-ROM, NIST Speech Disc.

Glasberg, B.R. and B.C.J. Moore, 1990. Derivation of auditory filter shapes from notched-noise data. Hear. Res., 47: 103-138. DOI: 10.1016/0378-5955(90)90170-T

Haton, J.P., C. Cerisara, D. Fohr, Y. Laprie and K. Smaïli, 2006. Reconnaissance automatique de la parole-Du signal à son interprétation: Du signal à son interprétation. Dunod.

Hermansky, H., 1990. Perceptual Linear Predictive (PLP) analysis of speech. J. Acoust. Soc. Am., 87: 1738-1752. DOI: 10.1121/1.399423

Hirsch, H. and D. Pearce, 2000. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. Proceedings of the Automatic Speech Recognition: Challenges for the new Millenium, Sept. 18-20, Paris, France.

Irino, T. and R.D. Patterson, 1997. A time-domain, level-dependent auditory filter: The gammachirp. J. Acoust. Soc. Am., 101: 412-419. DOI: 10.1121/1.417975

Irino, T. and R.D. Patterson, 2006. A dynamic compressive gammachirp auditory filterbank. IEEE Trans. Audio Speech Lang. Process., 14: 2222-2232. DOI: 10.1109/TASL.2006.874669

Martens, J.P. and L. Van Immerseel, 1990. An auditory model based on the analysis of envelope patterns. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Apr. 3-6, IEEE Xplore Press, Albuquerque, NM., pp: 401-404. DOI: 10.1109/ICASSP.1990.115713

Meddis, R., E.A. Lopez-Poveda, R.R. Fay and A.N. Popper, 2010. Computational Models of the Auditory System. 1st Edn., Springer, New York, ISBN-10: 1441959343, pp: 350.

Missaoui, I. and Z. Lachiri, 2014. Gabor filterbank features for robust speech recognition. Proceedings of the International Conference on Image and Signal Processing, Jun. 30-Jul. 2, Springer, France, pp: 665-671. DOI: 10.1007/978-3-319-07998-1_76

Moore, B.C.J., 2012. An Introduction to the Psychology of Hearing. 6th Edn., BRILL, Bingley, ISBN-10: 1780520387, pp: 441.

Patterson, R., I. Nimmo-Smith, J. Holdsworth and P. Rice, 1987. An efficient auditory filterbank based on the gammatone function. Proceedings of the Meeting of the IOC Speech Group on Auditory Modelling at RSRE, (GAM' 87).

Patterson, R.D., M. Unoki and T. Irino, 2003. Extending the domain of center frequencies for the compressive gammachirp auditory filter. J. Acoust. Soc. Am., 114: 1529-1542. DOI: 10.1121/1.1600720

Perdigao, F. and L. Sá, 1998. Auditory models as front-ends for speech recognition. Proceeding of the NATO ASI on Computational Hearing, (CH' 98), pp: 179-184.

Rabiner, L. and B.H. Juang, 1993. Fundamentals of Speech Recognition. 1st Edn., PTR Prentice Hall, Englewood Cliffs, ISBN-10: 0130151572, pp: 507.

Shao, Y., Z. Jin, D. Wang and S. Srinivasan, 2009. An auditory-based feature for robust speech recognition. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Apr. 19-24, IEEE Xplore Press, Taipei, pp: 4625-4628. DOI: 10.1109/ICASSP.2009.4960661

Unoki, M., T. Irino, B. Glasberg, B.C. Moore and R.D. Patterson, 2006. Comparison of the roex and gammachirp filters as representations of the auditory filter. J. Acoust. Soc. Am., 120: 1474-1492. DOI: 10.1121/1.2228539

Van Immerseel, L.M. and J.P. Martens, 1992. Pitch and voiced/unvoiced determination with an auditory model. J. Acoust. Soc. Am., 91: 3511-3526. DOI: 10.1121/1.402840

Wang, D.L. and G.J. Brown, 2006. Computational Auditory Scene Analysis: Principles, Algorithms and Applications. 1st Edn., Wiley, Hoboken, ISBN-10: 0471741094, pp: 395.

Young, S., G. Evermann, M. Gales, T. Hain and D. Kershaw *et al*., 2009. The HTK book version 3.4.1. Cambridge University Engineering Department, Cambridge, U.K.

Zouhir, Y. and K. Ouni, 2013. Speech Signals Parameterization Based on Auditory Filter Modelling. In: Advances in Nonlinear Speech Processing, Drugman, T. and T. Dutoit (Eds.), Springer, Heidelberg, ISBN-10: 3642388477, pp: 60-66.

Zouhir, Y. and K. Ouni, 2014. A bio-inspired feature extraction for robust speech recognition. SpringerPlus, 3: 651-651. DOI: 10.1186/2193-1801-3-651

Zouhir, Y. and K. Ouni, 2015. Noise robust speech parameterization using relative spectra and auditory filterbank. Res. J. Applied Sci. Eng. Technol., 9: 755-759.