International Journal of Electrical and Electronics Engineers
Vol. No. 9, Issue No. 02, July - December 2017
ISSN (O) 2321-2055
ISSN (P) 2321-2045

Speech Recognition using PNCC

Prachi P. Dani[1], Mrs. M.S. Deole[2]

[1]E&TC Department, R.H. Sapat College of Engineering,
Mgmt Studies and Research, Nashik, (India)

ABSTRACT

There are several approaches of feature extraction algorithms in speech recognition, e.g. Mel frequency cepstral coefficients (MFCC) [1], perceptual linear prediction (PLP) [2] and power-normalized cepstral coefficients (PNCC) [3]. PNCC a new feature extraction algorithm based on auditory processing is described in this paper. The new features of PNCC processing include the use of a power-law nonlinearity that has been replaced by the traditional log nonlinearity used in MFCC coefficients. There is use of medium-time power analysis, in which environmental parameters are estimated over a longer duration than is commonly used for speech, as well as frequency smoothing. PNCC is basically used for the improvement in recognition accuracy in noisy conditions. Paper shows the Features extracted using PNCC and impoved recognition accuracy using PNCC algorithm.The related results are checked based on subjective measure. In subjective measure SNR, Peak Signal-to-Noise Ratio (PSNR), Segmental-SNR and Mean Square Error (MSE),Root mean square error(RMSE)are assumed.

Keywords: ***Speech recognition, feature extraction, Mel frequency cepstral coefficients, Signal to noise ratio, Peak signal to noise ratio, Mean square error, automatic speech recognition.***

I. INTRODUCTION

Nowadays the performance of speech recognition systems in acoustical environments has drastically improved. Most speech recognition systems remain sensitive to the nature of the and their performance decreases sharply in the presence of sources of degradation such as additive noise, linear channel distortion, and reverberation. According to the speech recognition process, recognition technology can be divided into four classifications. Noise reduction in time-frequency domain, such as spectral subtraction and wiener filter, is the earliest technology. To compensate the noise in feature level is the second classification. Vector Taylor series (VTS) [1] is a popular method in this category. The third classification is noise compensation in model level. One of the most challenging problem is that recognition accuracy degrades significantly if the test environment is different from the training environment and if the acoustical environment includes disturbances such as additive noise, channel distortion, speaker differences, reverberation.

The presently developed systems for automatic speech recognition are based on two types of features mel frequency cepstral coefficients (MFCC) [2] and perceptual linear prediction (PLP) coefficients [4] Spectro-temporal has been observed that two-dimensional Gabor filters provide a reasonable approximation to the spectro temporal response, which has leads to various approaches to extract features for speech recognition.[4] This paper describe the brief introduction of an additionalfeature set for speech recognition referred as power normalized cepstral coefficients (PNCC) and implementation of MFCC speech recognition. Mel Frequency cepstral Coefficients (MFCC) is a widely used feature extraction method implemented in multiple ways. Here

MFCC for speech recognition system is tested using Matlab 2014 software,which is also used in therecognition tests.The development of PNCC feature extraction was motivated by a desire to obtain a set of practical features for speech recognition that are more robust with respect to acoustical variability in their native form, without loss of performance when the speech signal is undistorted, and with a degree of computational complexity that is comparable to that of MFCC and PLP coefficients. While many of the attributes of PNCC processing have been strongly influenced by consideration of various attributes of human auditory processing. There is one approach that provides pragmatic gains in robustness at small computational cost over approaches that are more faithful to auditory physiology in developing the specific processing that is performed.

II. LITERATURE SURVEY

Chanwoo Kim and Richard M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) For Robust Speech Recognition"[1] states the new feature extraction algorithm called Power Normalized Cepstral Coefficients (PNCC) that is motivated by auditory processing. Major new features of PNCC processing include the use of a power-law nonlinearity that replaces the traditional log nonlinearity used in MFCC coefficients, a noise-suppression algorithm based on asymmetric filtering that suppresses background excitation, and a module that accomplishes temporal masking.

Kuansan Wang and Shihab Shamma, "Self-Normalization and Noise-Robustness in Early Auditory Representations," [2] address the contribution of operations to the formation of robust and perceptually significant representation in the early auditory system. The auditory representation of the acoustic spectrum is effectively a self-normalized spectral analysis, i.e. the auditory system computes a spectrum divided by a smoothed version of itself. Such a self-normalization induces significant such as spectral shape enhancement and robustness against scaling and noise corruption.

P. J. Moreno, B. Raj, and R. M. Stern,[3] presents the use of a Vector Taylor series(VTS) expansion to characterize efficiently and accurately the impacts on speech statistics of unknown additive noise and unknown linear filtering in a transmission channel. The VTS approach is computationally efficient. It can be applied either to the incoming speech feature vectors, or to the statistics representing these vectors. In the first case the speech is compensated and then recognized in the second case HMM statistics are modified using the VTS formulation. Both approaches use only the actual speech segment being recognized to compute the parameters required for environmental compensation.

Li Deng, JashaDroppo, and Alex Acero, "Estimating Cepstrum of Speech Underthe Presence of Noise Using a Joint Prior of Static and Dynamic Features,"[6] presents a new algorithm for statistical speech feature enhancement in the cepstral domain. The algorithm exploits joint prior distributions in the clean speech model, which incorporate both the static and frame-differential dynamic cepstral parameters. Clean speech given the noisy observation are computed using a linearized version of a nonlinear acoustic distortion model, and, based on this linear approximation, the conditional minimum mean square error (MMSE) estimator for the clean speech feature is derived.

P. Pujol, D. Macho, and C. Nadeu [7] focuses on the recursive updating of MVNparameters, paying attention to the involved algorithmical delay. First, there is a decoupling of the look-ahead factor and the initial estimation of mean and variance, and latter there is a key factor for the recognition performance. Then, several kinds of

initial estimations that make sense in different application environments are tested, and their performance is compared.

Teddy Surya Gunawan and Eliatham by Ambikairajah, "A new forward maskingmodel and its application to speech enhancement" [8] States a new forward masking model, which is applied to speech enhancement. The model develops a novel expression for forward masking, where the parameters are related to the masker level, the delay and the frequency obtained by curve cutting the psychoacoustic data. It Provides significant improvements over existing speech enhancement methods, when tested with speech signals corrupted by various noises at very low signal to noise ratios.

A. Schwarz, B. Mertsching M. Brucke, W. Nebel [9] states that application of themodel is used to determine optimized word lengths ina hardware. The development of the perception model as a FPGA/ASIC for a target system, provides efficient co-processing power and allows real time implementations of complex auditory-based speech processing algorithms.

Samuel Thomas, Sriram Ganapathy,and Hynek Hermansky," Recognition of Reverberant Speech Using Frequency Domain Linear Prediction"[11] presents a feature extraction technique based on modeling temporal envelopes of the speech signal in narrow sub-bands using frequency domain linear prediction (FDLP). FDLP provides an all-pole approximation

III. STRUCTURE OF PNCC ALGORITHM

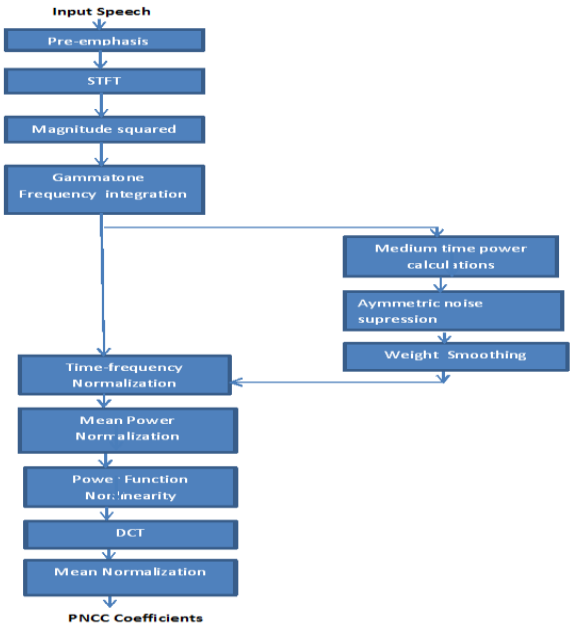


Fig.1 Block diagram of PNCC Algorithm.

Figure 1 shows the structure of conventional PNCCprocessing concepton the basis of recognition accuracy, which is introduce in this paper. The major innovations of PNCC processing include the redesigned nonlinear rate-intensity function, along with the series of processing elements to suppress the effects of background acoustical activity based on medium-time analysis. As in figure, the initial processing stages of PNCC processing are similar to the corresponding stages of MFCC and PLP analysis, except that the frequency analysis is performed using gammatone filters. This is followed by the series of nonlinear time-varying

operations that are performed using the longer-duration temporal analysis that accomplish noise subtraction as well as a degree of robustness with respect to reverberation. The final stages of processing are also similar to MFCC and PLP processing, with the exception of the carefully-chosen power-law nonlinearity with higher exponent.

III. COMPONENTS OF PNCC PROCESSING

1.1 Initial Processing

In MFCC features, a pre-emphasis filter of the form $H(z) = 1 - 0.97z{-}1$ is applied. A short-time Fourier transform (STFT) is performed using Hamming windows of duration 25.6 ms, with 10 msbetween frames, using a DFT size of 1024. Spectral power in 40 analysis bands is obtained by weighting the magnitude-squared STFT outputs for positive frequencies by the frequency response associated with a 40-channel gammatone-shaped filter bank whose center frequencies are linearly spaced in Equivalent Rectangular Bandwidth (ERB)[5] between 200 Hz and 8000 Hz, using the implementation of gammatone filters.

1.2 Temporal Integration for Environmental Analysis

Most speech recognition and speech committal to writing systems use analysis frames of length between twentymsand thirty ms. it's oftentimes ascertained that longer analysis windows give higher performance for noise modelling and environmental normalization [6] as a result of the facility related to most background conditions changes additionalslowly than the instant power related to speech. additionally, Hermanskyetal have ascertained that the characterization and exploitation information concerning the longer-term envelopes of every gamma tone channel will givecomplementary information that's helpful for rising speech recognition accuracy, In PNCC processing there is an estimate a quantity that is referred to as"medium-time power" $Q[m, l]$ by computing the running average of $P[m, l]$, the power observed in a single analysis frame, according to the equation:

$$\tilde{Q}[m,l] = \frac{1}{2M+1} \sum_{m'=m-M}^{m+M} P[m',l]$$

Where m represents the frame index and l is the channel index.

1.3 Asymmetric Noise Suppression

This gives a new approach to noise compensation which is referred to as asymmetric noise suppression (ANS). The concept is that the speech power in each channel usually changes more rapidly than the background noise power in the same channel. or it is said that speech usually has a higher-frequency modulation spectrum than noise.Nowadays many algorithms, RASTA-PLP processing, have been developed using either high-pass filtering or band-pass filtering in the modulation spectrum domain either explicitly or implicitly. The simplestway to accomplish this objective is to perform high-pass filtering in each channel which has the effect of removing slowly-varying components which typically represent the effects of additive noise sources rather than the speech signal. One common problem with the application concept of conventional linear high-pass filtering in the power domain is that the filter output can become negative. Negative values for the power coefficients can cause problems in the application of the compressive nonlinearity and in speech resynthesis unless a suitable floor value is applied to the power coefficients [7]. Rather than filtering in the power domain,

performing filtering after applying the logarithmic nonlinearity, as is done with conventional cepstral mean normalization in MFCC processing. Spectral subtraction is another way to reduce the effects of noise, whose power changes slowly. In spectral subtraction techniques, the noise level is typically estimated from the power of non-speech segments (*e.g.* [7])

1.4 Temporal Masking

Many algorithms showed that the human auditory system appears to focus more on the onset of an incoming power envelope rather than the falling edge of that same power envelope. This has led to several onsets enhancement algorithms and shown in fig.2. There is one way to overcome this effect in PNCC processing, that is done by obtaining a moving peak for each frequency channel *l* and suppressing the instantaneous power if it falls below this envelope. The processing invoked for temporal masking is given in block diagram. Peak power *Qp*[*m, l*] for each channel using the following equation:

$$\tilde{Q}_p[m,l] = \max\left(\lambda_t \tilde{Q}_p[m-1,l], \tilde{Q}_0[m,l]\right)$$

Where λt is the forgetting factor for obtaining the online peak. Where m is the frame index and *l* is the channel index.

1.5 Rate Level Nonlinearity

There is a critical importance of the nonlinearfunction that describes the relationship between incoming signal amplitude in a given frequency channel and the corresponding response of the processing model. This "rate-level nonlinearity" is explicitly or implicitly a crucial part of every model of auditory processing.

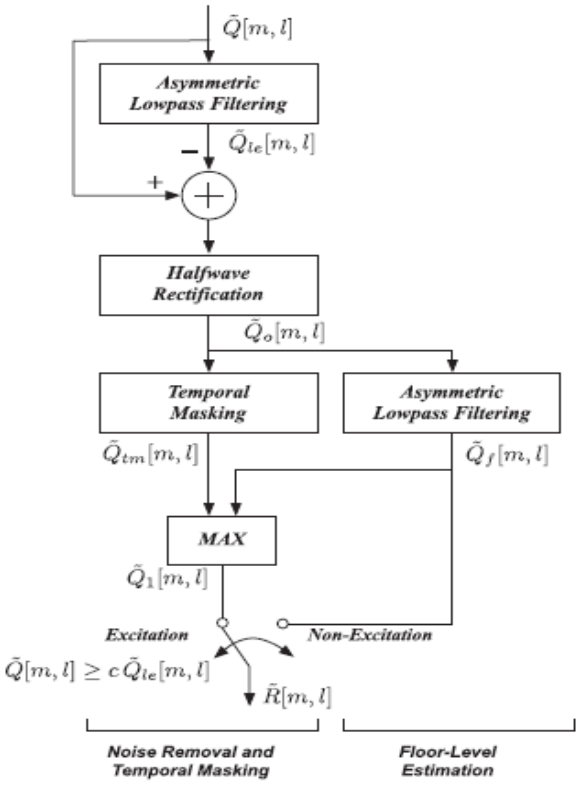


Fig.2. Functional block diagram of the modules for asymmetric noise suppression (ANS) and temporal masking in PNCC processing.

International Journal of Electrical and Electronics Engineers
Vol. No. 9, Issue No. 02, July - December 2017
ISSN (O) 2321-2055
ISSN (P) 2321-2045

IV. SUBJECTIVE MEASURES

4.1 Signal to noise ratio (SNR)

Signal-to-noise ratio is a term used in field of engineering that compares the level of a desired signal to the level of background noise. It is defined as the ratio of signal power to the noise power, and expressed in decibels. A ratio higher than 1:1 indicates more signal than noise. While SNR is commonly used for speech signals. It is represented by following equation.

$$\mathrm{SNR} = \frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}.$$

4.2 Peak signal to noise ratio (PSNR)

Peak signal-to-noise ratio, is an engineering term for the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. Because many signals have a very wide dynamic range, PSNR is usually expressed in terms of the logarithmic decibel scale.

$$\mathbf{PSNR = 20.\ log_{10}(MAX_I) - 10.log_{10}(MSE)}$$

4.3 Mean Square Error (MSE)

The MSE is a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better. In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors or deviations that is, the difference between the estimator and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss or quadratic loss.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

4.4 Root-mean-square error (RMSE)

It is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed. The RMSD represents the sample standard deviation of the differences between predicted values and observed values. These individual differences are called residuals when the calculations are performed over the data sample that was used for estimation, and are called prediction errors when computed out-of-sample. The RMSD serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power. RMSD is a measure of accuracy, to compare forecasting errors of different models for a particular data and not between datasets, as it is scale-dependent.

$$\mathrm{RMSD}(\hat{\theta}) = \sqrt{\mathrm{MSE}(\hat{\theta})} = \sqrt{\mathrm{E}((\hat{\theta} - \theta)^2)}.$$

International Journal of Electrical and Electronics Engineers
Vol. No. 9, Issue No. 02, July - December 2017
ISSN (O) 2321-2055
ISSN (P) 2321-2045

V RESULTS AND DISCUSSIONS

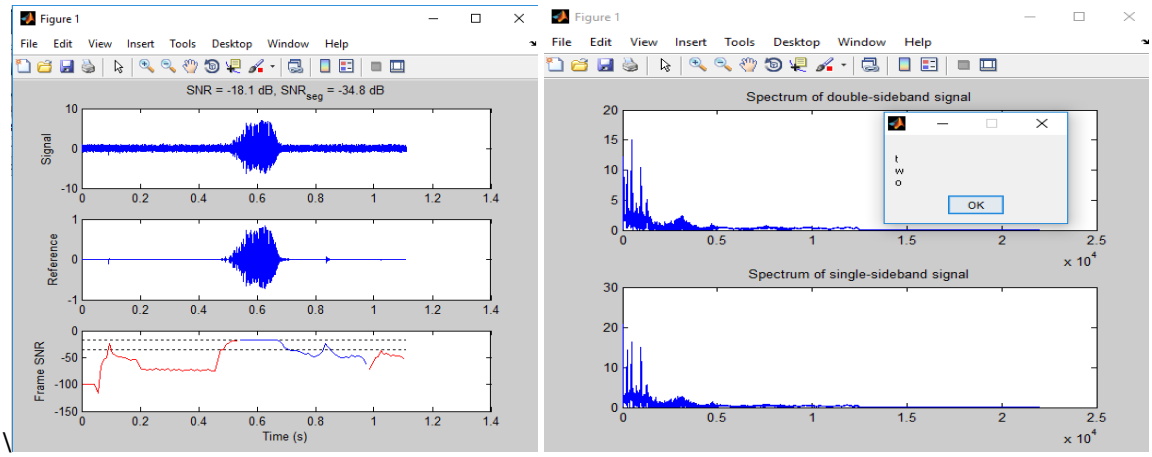
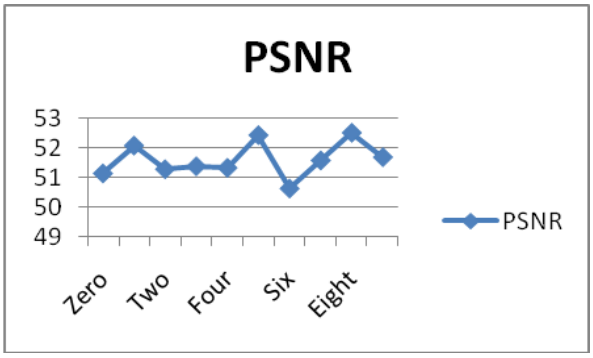
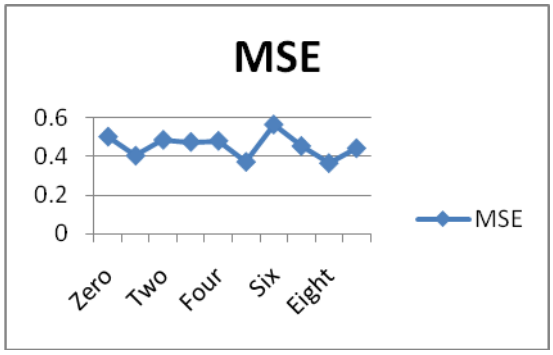


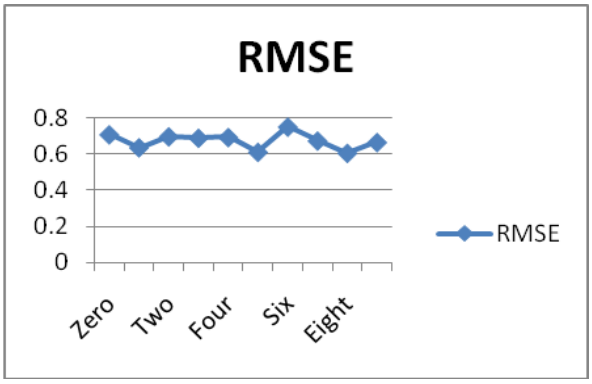
Fig 3: PNCC Output waveforms



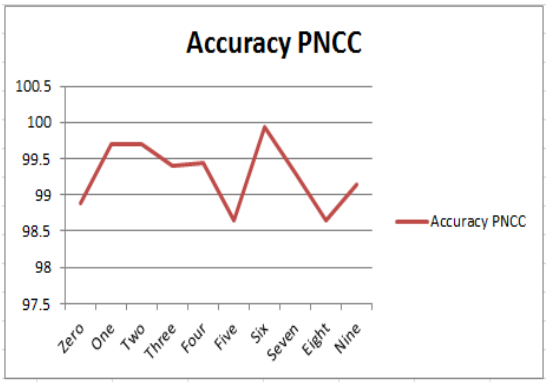
Fig(a) Graphical representation of PSNR



Fig(b) Graphical representation of MSE



Fig(c) Graphical representation ofRMSE



Fig(d) Recognition accuracy of PNCC

VI. CONCLUSION

This paper presents an extraction algorithm called PNCC (Power Normalized cepstral coefficients). Currently, many new schemes are proposed in the field of speech recognition. So the best method among all should be found out. The proposed method is among the efficient method of all to noise removal which leads to extract features for speech recognition. Many techniques are proposed for automatic speech recognition but none of it is considered to be perfect for measurement of accuracy. Improving accuracy plays a crucial role in the field of

speech processing. In this paper features are extracted using PNCC with real time as well as from standard database and accuracy has been checked using PNCC (Power Normalized cepstral coefficients) algorithm. From the estimated results it is found that PNCC algorithm reduces the computational complexity with improved accuracy. Logically, a bigger value of SNR is good because it means that the ratio of signal to noise is higher. Higher SNR indicate that higher removal of noise. The related results are checked based on subjective measure. In subjective measure SNR, Peak Signal-to-Noise Ratio (PSNR), Segmental-SNR and Mean Square Error (MSE), Root mean square error(RMSE)are assumed.

VI. ACKNOWLEDGEMENTS

First and the foremost I, take this opportunity to express gratitude to my guide, Mrs. M.S. Deole, for her constant encouragement and support throughout the project implementation. I sincerely thank Prof. S. P. Agnihotri, Head of Department of Electronics & Telecommunication Engineering for his advice and support during course of this work.With deepsense of gratitude I thank to our PrincipalDr.P. C. Kulkarni and Management of Gokhale Education Society for providing all necessary facilities and their constant encouragement and support. I also express my thanks to all teaching and non-teaching staff for their kind co-operation and guidance also.

REFERENCES

[1] Chanwoo Kim and Richard M. Stern, “Power-Normalized Cepstral Coefficients (PNCC) For Robust Speech Recognition" IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 2016.

[2] Kuansan Wang and Shihab Shamma, “Self-Normalization and Noise-Robustness in Early Auditory Representations," IEEE TRANSACTIONS ON SPEECH ANDAUDIO PROCESSING, VOL. 2, NO. 3, JULY 1994.

[3] P. J. Moreno, B. Raj, and R. M. Stern, “A vector Taylor series approach forenvironment - independent speech recognition, “IEEE Int. Conf. Acoust, Speech and Signal Processing, May. 1996, pp. 733{736}.

[4] Steven .B. Davis, “Comparison of Parametric representations for monosyllabicword recognition in continuously spoken sentences, " Proceedings of the IEEE, vol. 81, No. 9, pages 1215-1247, 1993.

[5] M. G. Heinz, X. Zhang, I. C. Bruce, and L. H. Carney,\Auditory nerve modelfor predicting performance limits of normal and impaired listeners," Acoustics Research Letters Online, vol. 2, no. 3, pp. 91{96, July 2001}.

[6] Li Deng, Jasha Droppo, and Alex Acero, “Estimating Cepstrum of Speech Underthe Presence of Noise Using a Joint Prior of Static and Dynamic Features, " IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 12, NO. 3, MAY 2004.

[7] P. Pujol, D. Macho, and C. Nadeu, “On real-time mean-and-variance normalizationof speech recognition features," IEEE Int. Conf. Acoustic Speech and Signal Processing,vol. 1, May 2006, pp. 773{776}.

[8] Teddy Surya Gunawan and Eliatham by Ambikairajah, “A new forward maskingmodel and its application to speech enhancement, " ICASSP 2006, 142440469X/06/20.00 c 2006 IEEE.

[9] Schwarz, B. Mertsching M. Brucke, W. Nebel, "Implementing a Quantitative Model for the Effective Signal Processing in the Auditory System on a Dedicated Digital VLSI Hardware."

[10] N. Mesgarani, M. Slaney, and S. Shamma,\Discrimination of speech from nonspeechbased on multiscale spectro-temporal modulations, " IEEE Trans. Audio, Speech, and Language Processing, vol. 14, no. 3, pp. 920{929, May 2006}.

[11] Samuel Thomas, Sriram Ganapathy, and Hynek Hermansky, "Recognition of Reverberant Speech Using Frequency Domain Linear Prediction," ." IEEE SIGNALPROCESSING LETTERS, VOL. 15, 2008.

[12] S. R. Mahadeva Prasanna, B. V. Sandeep Reddy, and P. Krishnamoorthy, "VowelOnset Point Detection Using Source, Spectral Peaks, and Modulation Spectrum Energies," IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 17, NO. 4, MAY 2009.

[13] C. Kim and R. M. Stern, "Power function-based power distribution normalizationalgorithm for robust speech recognition," IEEE Automatic Speech Recognition and Understanding Workshop, Dec. 2009, pp. 188{193}.

[14] Adrian Pass Ji, Ming Philip Hanna Jianguo, Zhang Darryl Stewart, "Inter-FrameContextual Modelling For Visual Speech Recognition," IEEE 17th International Conference on Image Processing September 26-29, 2010, Hong Kong.

[15] C. Kim, K. Kumar, and R. M. Stern, "Binaural sound source separation motivatedby auditory processing," IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, May 2011, pp. 5072{5075}.

[16] Abigail A. Kressner, David V. Anderson, and Christopher J. Rozell, "Robustness Of The Hearing Aid Speech Quality Index (Hasqi)," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics October 16-19, 2011, New Paltz, NY.

[17] F. Muller and A. Mertins, "Contextual invariant-integration features for improvedspeaker-independent speech recognition," Speech Communication, vol. 53, no. 6, pp. 830{841, July 2011.

[18] Tianyu T. Wang and Thomas F. Quatieri, "Two-Dimensional Speech-Signal Modeling, "IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO.6,AUGUST 2012.

[19] Niko Moritz, JornAnem□uller and Birger Kollmeier, "An Auditory Inspired AmplitudeModulation Filter Bank for Robust Feature Extraction in Automatic Speech Recognition."IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 23, NO. 11, NOVEMBER 2015.