AM-FM Based Robust Speaker Identification in Babble Noise

Mangesh S. Deshpande
Department of E&TC Engineering
SRES College of Engineering,
Kopargaon, India

Raghunath S. Holambe
Department of Instrumentation Engineering
SGGS Institute of Engineering and Technology,
Nanded, India

ABSTRACT
Speech babble is one of the most challenging noise interference due to its speaker/speech like characteristics for speech and speaker recognition systems. Performance of such systems strongly degrades in the presence of background noise, like the babble noise. Existing techniques solve this problem by additional processing of speech signal to remove noise. In contrast to existing works, the aim is to improve noise robustness focusing on the features only. To derive robust features, amplitude modulation - frequency modulation (AM-FM) based speaker model is proposed. The robust features are derived by fusing the characteristics of speech production and speech perception mechanisms. The performance is evaluated using clean speech corpus from TIMIT database combined with babble noise from the NOISEX-92 database. Experimental results show that the proposed features significantly improve the performance over the conventional Mel frequency cepstral coefficient (MFCC) features under mismatched training and testing environments.

General Terms
Pattern Recognition, Algorithms, Experimentation.

Keywords
Speaker identification, AM-FM model, Babble noise.

1. INTRODUCTION
Speaker recognition is an important and emerging field with many potential applications such as voiced internet applications, telephone banking, security, surveillance etc [17]. Current speaker identification systems work well within controlled environments that are relatively noise-free. However, robustness of a speaker recognition system to additive background noise is an important problem when the system needs to operate in noisy environments [6, 10, 19].

This is an even more challenging task when the system has to perform recognition in a noisy environment different from that of training especially at low signal-to-noise ratio (SNR). This mismatch between training and testing environments seriously degrades the system performance. Various sources give rise to this mismatch, such as additive noise, channel distortion, different speaker characteristics, different speaking modes etc. Among all different additive noise signals (car noise, traffic noise, office noise, factory noise, babble noise etc.), one of the most challenging noise conditions is multi-speaker or babble noise environment, where the interference is speech from speakers in the vicinity. This noise is uniquely challenging because of it's highly time evolving structure and its similarity to the desired target speech [12].

A variety of techniques have been developed to improve the system performance in noisy environment [6, 8, 9, 10, 25]. For the purpose of handling additive noise, the techniques can be roughly categorized into two classes. The first class is model-based and the second class is feature-based. In the first class, some linear and nonlinear compensation techniques are used, so that the modified recognition models will be able to classify the mismatched testing speech features collected in the testing environment. The typical examples of this class include the well-known noise masking [8], speech and noise decomposition (SND) [32], hypothesized Wiener filtering [25], vector Taylor series (VTS) [1], maximum likelihood linear regression (MLLR), model-based stochastic matching [13], statistical reestimation (STAR), parallel model combination (PMC) [5] optimal subband likelihood weighting based on the criteria of minimum classification error (MCE) and maximum mutual information (MMI) [31] etc. On the other hand, within the feature-based approaches in the second class there are two subgroups. The first subgroup of approaches tries to modify the testing speech features and make them match the acoustic conditions better for the trained models. The well known spectral subtraction (SS), fixed codeword-dependent cepstral normalization (FCDCN), feature-based stochastic matching [13], multivariate Gaussian based cepstral normalization (RATZ) and MMSE estimation of clean speech by considering the phase relationship of speech and noise are typical examples of this subgroup. In the second subgroup of feature-based approaches, on the other hand, a special robust speech feature representation is developed to reduce the sensitivity to the various acoustic conditions and this feature representation is used for both training and testing. This paper covers this second approach.

Many existing speaker recognition systems use the short-time spectral information of a speaker's voice extracted in the form of a time-series of feature vectors, usually composed of linear prediction cepstral coefficients (LPCCs) or Mel frequency

cepstral coefficients (MFCCs). These features are based on the linear source-filter model of speech production [24, 26]. Systems based on these features have been shown to achieve remarkable performance in controlled conditions [20]. However, strong degradation of performance occurs in the presence of significant background noise and/or strong channel distortions. Therefore real-world robustness still appears to be an open research issue for speaker recognition systems.

In [2], preliminary results have indicated that a significant part of the acoustic information cannot be modeled by the linear source-filter acoustic model and thus, the need for nonlinear features becomes apparent. These features, which are based on either the FM or the AM part, provide additional acoustic information. These features can model the dynamic nature of speech and capture some of its fine structure and its rapid fluctuations. Furthermore, they appear to be relatively noise resistant and, thus, yield improved results, especially when a mismatch in the training and testing conditions is present. In order to tackle the problem of robust speaker identification in babble noise, in this paper, the use of nonlinear model parameters (AM-FM) as features is proposed. These features are based on frequencies of the speech signal derived from its phase is proposed. These features are effective when the testing and training data are recorded under different noise levels.

The rest of the paper is arranged as follows. The AM-FM speaker model along with multiband filtering and demodulation is described in Section 2. The feature extraction process is explained in Section 3. The Gaussian mixture modeling technique is shortly discussed in Section 4. The experiments conducted and the results obtained are mentioned in Section 5. Finally conclusions drawn are mentioned in Section 6.

2. AM-FM MODEL

The AM-FM modeling technique has been applied to speech signal analysis with varying degrees of success, in areas such as formant tracking [22], speech synthesis [14], speech recognition [3, 23] and speaker identification [11]. It was first proposed by Potamianos et al. [22] in the context of formant tracking. Marco Grimaldi et. al. [7] extended the same work to the problem of speaker identification. They indicated that the characterization of the different instantaneous frequencies within the speech signal play a significant role in capturing the identity of a speaker. Jankowski et al. [11] adopted the AM-FM model to characterize some fine structures of the human voice for the purpose of speaker identification. The AM-FM modeling technique can be effectively used for modeling the speech production system. Vocal tract resonances can change rapidly both in frequency and amplitude even within a single pitch period. This may be due to rapidly varying and separated speech airflow in the vocal tract [16]. The effective air masses in vocal tract cavities and effective cross sectional areas of the airflow vary rapidly, causing modulations of air pressure and volume velocity. This leads to the actual speech signal, $s(t)$ composed of a sum of N resonances as,

$$s(t)=\sum_{i=1}^{N} R_i(t) \tag{1}$$

where $R(t)$ is a single speech resonance, which can be represented as an AM-FM signal,

$$R(t)=a(t)\cos[2\pi(f_c t+\int_0^t q(\tau)d\tau)+\theta] \tag{2}$$

where f_c is the center value of the resonance (formant) frequency, $q(t)$ is the frequency modulating signal and $a(t)$ is the time varying amplitude. The individual resonance may be isolated by band-pass filtering the speech signal. The instantaneous resonance frequency signal is defined as,

$$f_i(t)= f_c + q(t). \tag{3}$$

The estimation of the amplitude envelope and instantaneous frequency components, i.e., the demodulation of each resonant signal, can be done with the energy separation algorithm (ESA), or utilizing the Hilbert transform demodulation (HTD) algorithm.

2.1 Hilbert Transform Demodulation

Numerous techniques have been proposed in the literature to perform the demodulation [16, 21, 4]. Although the digital energy separation algorithm (DESA) [22, 16] is computationally less expensive, the Hilbert transform demodulation (HTD) can give smaller error and smoother frequency estimates [16, 22]. In order to characterize a (single) instantaneous frequency for a real-valued signal, an analytic signal is first constructed; it is a transformation of the real signal into the complex domain. More formally, given a real input signal $s(t)$, its analytic signal $s_a(t)$ can be computed as

$$s_a(t)=s(t)+j\,\hat{s}(t), \tag{4}$$

where $\hat{s}(t)$ is the Hilbert transform of $s(t)$. We can decompose the analytic signal $s_a(t)$ as follows:

$$s_a(t)=a(t)e^{j\phi(t)}, \tag{5}$$

where

$$a(t)=|s_a(t)| \tag{6}$$

is called *instantaneous amplitude* (or Hilbert envelope) of the signal, and

$$\phi(t)=\angle s_a(t)=\arctan[\hat{s}(t)/s(t)] \quad (7)$$

is the *instantaneous phase*. The *instantaneous frequency* (IF) $f_i(t)$ is computed from the unwrapped *instantaneous phase* $\phi_u(t)$ as follows:

$$f(t)=\frac{1}{2\pi}\frac{d\phi_u(t)}{dt}. \quad (8)$$

The instantaneous frequency estimation is one of effective methods to detect and track frequency changes of a mono-component signal. But, in the case of multicomponent signals, the result becomes meaningless without breaking the signal down into its components [34].

2.2 Multiband Filtering and Demodulation

To obtain single resonance signal $R(t)$, from the speech signal $s(t)$, a filtering scheme can be used before demodulation, which is referred as multiband demodulation analysis (MDA). The MDA yields rich time-frequency information. MDA consists of a multiband filtering scheme and a demodulation algorithm. First, the speech signal is bandpass filtered using a filterbank, then each bandpass waveform is demodulated and its instantaneous amplitude and frequency are computed. The following steps are adopted to demodulate the speech signal and to extract the features:

• The speech signal $s(t)$ is bandpass filtered and a set of waveforms $w_k(t)$ is obtained (k denotes the output of the k th filter in the filterbank).

• For each bandpass waveform $w_k(t)$, its Hilbert transform $\hat{w}_k(t)$ is computed.

• The instantaneous amplitude, $a_{ik}(t)$ for each bandpass waveform is computed as,

$$a_{ik}(t)=\sqrt{w_k^2(t)+\hat{w}_k^2(t)}. \quad (9)$$

• The instantaneous frequency, $f_{ik}(t)$ for each bandpass waveform is computed as the first time derivative of the unwrapped phase $\phi_k(t)$ as,

$$f_{ik}(t)=\frac{1}{2\pi}.\frac{d\phi_k(t)}{dt}=\frac{1}{2\pi}.\frac{d}{dt}[\arctan(\hat{w}_k(t)/w_k(t))]. \quad (10)$$

After obtaining the instantaneous amplitude and frequency signals by demodulating each resonant signal, a short-time analysis is performed.

2.3 Short-Time Estimate: Frequency

Simple short-time estimate of the frequency F is the unweighted mean F_{iu} of the instantaneous frequency signal $f_i(t)$, i.e.,

$$F_{iu}=\frac{1}{\tau}\int_{t_0}^{t_0+\tau} f_i(t)dt, \quad (11)$$

where t_0 and τ are the start and duration of the analysis frame, respectively. Alternative estimate is the first weighted moment of $f_i(t)$ [22]. Using squared amplitude, $a_i^2(t)$ as the weight, the first weighted moment is,

$$F_{iw}=\frac{\int_{t_0}^{t_0+\tau}[f_i(t).a_i^2(t)]dt}{\int_{t_0}^{t_0+\tau}[a_i^2(t)]dt} \quad (12)$$

The adoption of a mean amplitude weighted instantaneous frequency is motivated by the fact that it provides more accurate frequency estimate and is more robust for low energy and noisy frequency bands when compared with an unweighted frequency mean [3,22].

To understand the behavior of F_{iu} and F_{iw}, let's consider the example given by Potamianos *et. al.* in [22]. Let the signal, $x(t)$ be a sum of two sinusoids with constant frequencies $f_1=1.5$ kHz, $f_2=1.7$ kHz and time-varying amplitudes $a_1(t)$ and $a_2(t)$:

$$x(t)=a_1(t)\cos[2\pi f_1 t]+a_2(t)\cos[2\pi f_2 t], \quad t\in[0,0.1]s. \quad (13)$$

where $a_1(t)=10t$ and $a_2(t)=1-10t$ therefore for the first half of the time interval (0 to 50 ms), the second sinusoid f_2 is dominant, while for the second half (50 to 100ms), f_1 dominates.

Figure 1 shows the amplitude envelope $|a(t)|$ of $x(t)$ and Figure 2 shows the instantaneous frequency $f_i(t)$ of $x(t)$ computed via HTD. At envelope maxima, the instantaneous frequency is equal to the average (amplitude weighted) frequency of the two sinusoids $f = (a_1 f_1 + a_2 f_2)/(a_1 + a_2)$, while at envelope minima, f presents spikes of value $f = (a_1 f_1 - a_2 f_2)/(a_1 - a_2)$; i.e., the spikes point toward the frequency of the sinusoid with the larger amplitude.

The short-time estimate F_{iu} and weighted estimate F_{iw} of the instantaneous frequency computed by HTD are shown in Figure 3. It shows that F_{iu} locks onto the sinusoid with the greater amplitude whereas F_{iw} provides a more 'natural' short-time estimate because the spikes of the instantaneous frequency correspond to amplitude minima and get weighted less in the F_{iw} average. Actually, F_{iw} is the mean weighted frequency of the two sinusoids, with squared amplitude as the weight.

These results can be generalized to the short-time frequency estimates of speech resonances by use of a sinusoidal speech model. A speech signal can be modeled as a sum of sinusoids with slowly time-varying amplitudes and frequencies [18]. In particular, a speech resonance can be modeled as a sum of a few sinusoids. The behavior of F_{iu} and F_{iw} estimates for a speech formant can then be viewed as a generalization of the two sinusoids case analyzed above. For a speech resonance signal, F_{iu} has the tendency to lock on the frequency with the greatest amplitude in the formant band, while F_{iw} weights each frequency in the formant band with its squared amplitude. Thus, the weighted frequency estimate F_{iw} provides more accurate formant frequencies and is more robust for low energy or noisy frequency bands [22].

3. FEATURE EXTRACTION

In the feature extraction process, the speech signal is first pre-emphasized using a pre-emphasis filter,

$$H(z) = 1 - 0.97z^{-1} \tag{14}$$

The pre-emphasized speech signal is then divided into 32 ms frames with 16 ms overlap and multiplied by Hamming window. The AM-FM features are computed from the instantaneous frequency and amplitude of each speech frame. In order to compute the instantaneous frequencies of the speech signal, a multiband demodulation analysis (MDA) is performed. It consists of a multiband filtering scheme and a demodulation algorithm. First, the speech signal is bandpass filtered with the

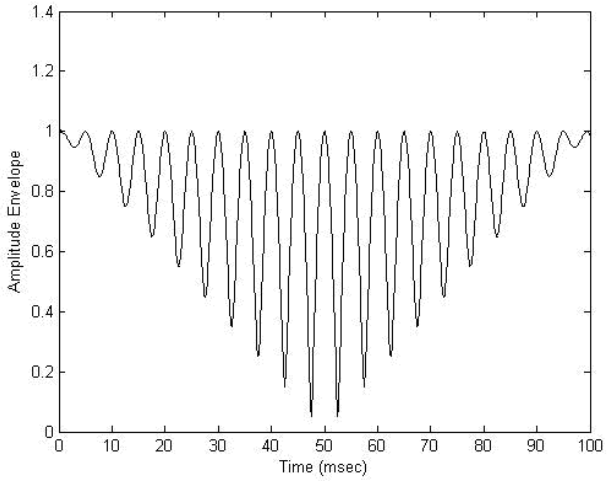

Fig 1: Amplitude envelope of $x(t)$ obtained using HTD

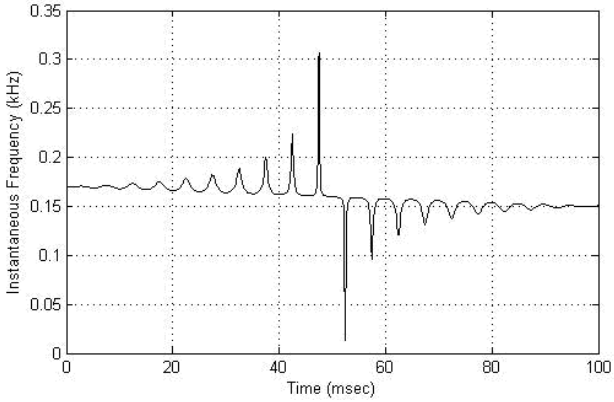

Fig 2: Instantaneous frequency of $x(t)$ obtained using HTD.

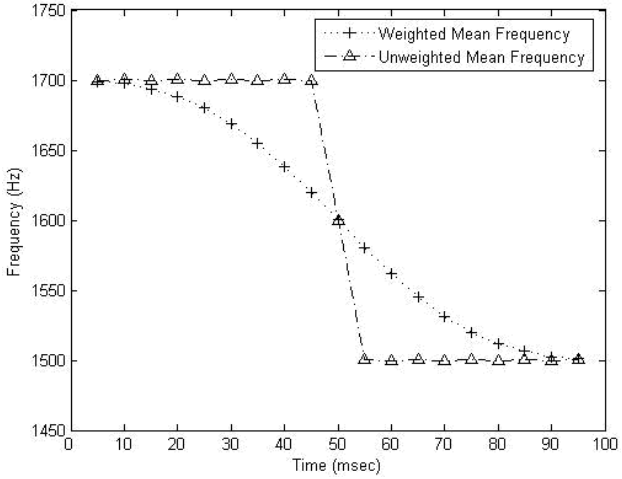

Fig 3: Weighted and unweighted estimates of the Instantaneous frequency of $x(t)$ obtained using HTD.

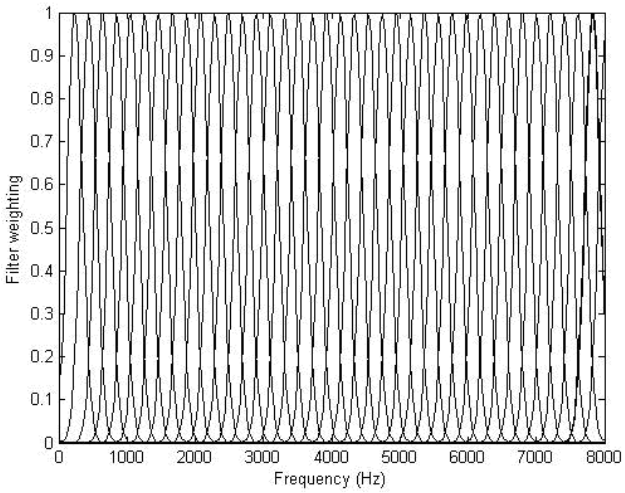

Fig 4: A filter-bank composed of 40 Gabor filters whose bandwidths are constant on the Hz scale.

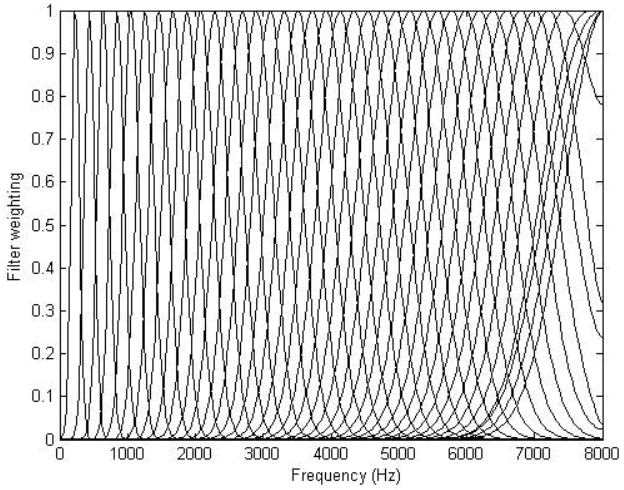

Fig 5: A filter-bank composed of 40 Gabor filters whose bandwidths are constant on the Mel scale.

use of a filterbank and then each bandpass waveform is demodulated using Hilbert transform demodulation and its instantaneous amplitude and frequency is computed. The filterbank used consists of a set of Gabor bandpass filters with center frequencies that are uniformly spaced on the frequency axis. Gabor filters are chosen because they are optimally compact and smooth in both the time and frequency domains. This characteristic guarantees accurate amplitude and frequency estimates in the demodulation stage [22] and reduces the incidence of ringing artifacts in the time domain [11]. Bandwidth of each individual Gabor filter within the filterbank plays an important role. Fine tuning the bandwidth of the filters used for speech analysis and speech characterization is a standard practice found in many approaches.

In this paper, two different filterbanks are used. The first filterbank (uniform) consists of 40 Gabor filters with uniformly spaced center frequencies and constant bandwidth of 200 Hz as shown in Figure 4. The second filterbank (non-uniform) consists of 40 Gabor filters which are non-uniformly spaced and the bandwidth varies according to the Mel scale as shown in Figure 5. This filterbank is very similar to the filterbank used in conventional MFCC feature extraction technique. The only difference is, instead of using triangular filters, Gabor filters are used. After obtaining the instantaneous amplitude and frequency using HTD, the short-time mean amplitude weighted instantaneous frequency estimate is obtained using Eq. (12). The estimate of short-time instantaneous frequency is expressed in kiloHertz in order to overcome the problem associated with the nodal variances of the Gaussian mixture model (GMM). Finally DCT is applied and only first 24 coefficients excluding zeroth coefficient are used to construct a feature vector. The feature vectors obtained using uniform filterbank are referred as F-1 and that of non-uniform filterbank as F-2.

4. SPEAKER MODELING USING GMM

Gaussian mixture modeling classifier is the most widely used probabilistic technique for speaker recognition [29, 27, 28, 30]. This classifier is able to approximate the distribution of the acoustic classes representing broad phonetic events occurring in speech production (e.g., during the production of vowels, nasals, fricatives etc.) and often outperforms other algorithms on the problem of speaker identification [33, 30].

A Gaussian mixture density is a weighted sum of M component densities and is given by the equation,

$$p\left(\vec{x}/\lambda\right) = \sum_{i=1}^{M} c_i p_i\left(\vec{x}\right), \tag{15}$$

where $\vec{x}$ is a D dimensional feature vector, c_i, $i=1,\ldots,M$ are the mixture weights and $p_i(\vec{x})$, $i=1,\ldots,M$, are the component densities of the form,

$$\frac{1}{2\pi^{D/2}\left|\sum_i\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\vec{x}-\vec{\mu}_i\right)'\sum_i{}^{-1}\left(\vec{x}-\vec{\mu}_i\right)\right\} \tag{16}$$

with mean vector $\vec{\mu}_i$ and covariance matrix $\sum_i$. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} c_i = 1$. The complete Gaussian mixture density is represented by the notation,

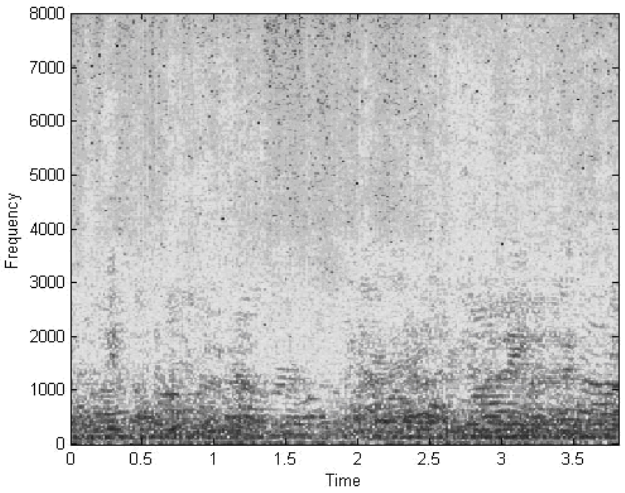

Fig 6: Spectrum of babble noise.

$$\lambda = \{ p_i, \vec{\mu}_i, \Sigma_i \}, \; i=1,\ldots,M. \quad (17)$$

Given training utterance of a speaker, the goal of speaker model training is to estimate the parameters of the GMM, λ using the *expectation-maximization* (EM) algorithm [30].

For speaker identification, a group of S speakers $S = \{1,2,3,\ldots..,s\}$ is represented by GMM's $\lambda_1, \lambda_2, \ldots, \lambda_s$. The objective is to find the speaker model which has the maximum a posteriori probability for a given observation sequence $X = \vec{x}_1, \ldots\ldots, \vec{x}_T$, for an utterance with T frames. The maximum a posteriori probability can be obtained by,

$$\hat{S} = \arg \max_{1 \le k \le s} \sum_{t=1}^{T} \log p \left(\vec{x}_t \middle/ \lambda_k \right), \quad (18)$$

in which $p(\vec{x}_t / \lambda_k)$ is given in Eq.(15).

5. EXPERIMENTS AND RESULTS

5.1 Database Description

Speaker identification experiments were carried out using TIMIT and NOISEX-92 databases. TIMIT is a noise free speech database recorded using a high quality microphone sampled at 16 kHz. It consists of 630 speakers, 70% male and 30% female from 8 different dialect regions in America. The speech is designed to have rich phonetic contents. It consists of 2 dialect sentences (SA), 450 phonetically compact sentences (SX) and 1890 phonetically diverse sentences (SI). Out of 10 spoken sentences per speaker, eight sentences, five SX and three SI (approximately 24 seconds) are used for training the speaker models. The two SA sentences (a total of 1260 tests of 3 seconds each) are used for testing and average identification results are noted. NOISEX-92 is a noise database which

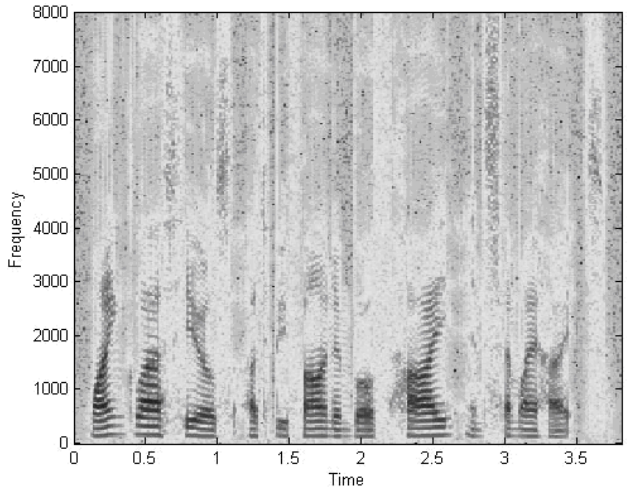

Fig 7: Spectrum of speech signal 'Wine glass heels are to be found in both high and semi-heights.

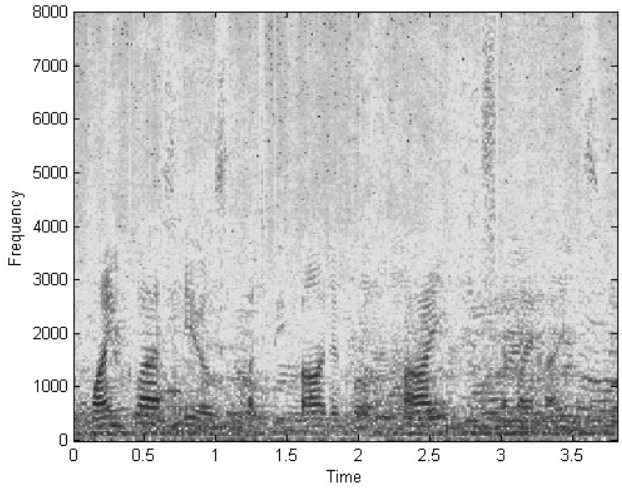

Fig 8: Spectrum of speech signal corrupted by babble noise with 0 dB SNR.

provides various noise signals recorded in real environments. For babble noise, the source of the babble is 100 people speaking in a canteen. The room radius is over two meters; therefore, individual voices are slightly audible.

5.2 Baseline Features

In order to compare the performance of the proposed features with that of most widely used MFCC features, MFCC features were extracted using a filterbank consisting of 40 triangular filters spaced between 0 and 8000 Hz. The feature vector consists of 24 cepstral coefficients excluding the zeroth coefficient.

5.3 Speaker Modeling

In the experiments performed, the speaker models are trained using 32 mixture (diagonal covariance) GMM. All models are trained using clean speech and the test dataset is corrupted by noise to obtain different SNRs.

5.4 Performance Evaluation

In order to evaluate the performance of the proposed features under mismatched conditions, the GMM speaker models (with 32 mixtures) were trained with clean speech only and babble noise was added to the test data to obtain SNR of 20, 10, 5 and 0 dB. Figure 6 shows the spectrum of babble noise. It shows that the characteristics of the babble noise are very similar to the speech signal, mostly covering the low frequency spectrum. Figure 7 shows the spectrum of the sentence, 'Wine glass heels are to be found in both high and semi-heights' spoken by a male speaker from the TIMIT database. It also covers mainly the low frequency spectrum. With the addition of babble noise, as the spectra overlaps, it is challenging to identify a speaker from the noisy speech. Figure 8 shows the spectrum of the speech signal corrupted by babble noise with 0 dB SNR.

Table 1 shows speaker identification rate in percent for the features F-1, F-2 and the MFCC under mismatched conditions. It shows that, speaker identification rate decreases with decreasing SNR. At higher SNR values, the F-2 features work equally well compared to the MFCC features and at low SNR values, the identification accuracy is better than the MFCC features. It confirms that, the MFCC features are well suited only when the training and testing speech is clean (noise free) and recorded in the same environment. Furthermore, MFCC takes into account only the speech perception mechanism and not the speech production mechanism. Features F-1 are based on speech production mechanism, whereas features F-2, considerers both speech production (using AM-FM approach) as well as perception (non-uniform filter bank) mechanism, hence more robust compared to MFCC features.

Table 1. Speaker identification performance obtained with the addition of babble noise in the test speech utterances at different SNRs using MFCC, F-1 and F-2 features.

Features	Speaker identification rate (%)			
	SNR= 20dB	**SNR= 10dB**	**SNR= 5dB**	**SNR= 0dB**
MFCC	96.25	77	49.25	18.5
F-1	89.75	79	59.25	28.5
F-2	96.75	88.25	66.5	37.75

6. CONCLUSIONS

Robustness of a speaker identification system to additive background noise is an important problem when the system needs to operate in noisy environments. Speech babble is one of the most challenging noise interference due to its speaker/speech like characteristics. To derive robust features, in this paper, an AM-FM based speaker model is proposed which combines the speech production and perception mechanism. These features show significant improvement in the speaker identification rate under mismatched training and testing environments compared to MFCC features. This paper shows that, instead of deriving the features based on only speech production or speech perception mechanism, if these two mechanisms are combined together, it is possible to obtain robust features, which shows further improvement in speaker identification accuracy.

7. REFERENCES

[1] Acero, A., Dend, L., Kristjansoon, T. and Zhang, J. 2000. Hmm adaptations using vector Taylor series for noisy speech recognition. In Proceedings of (ICSLP'00), 869-872.

[2] Dimitriadis, D. and Maragos, P. 2003. Robust energy demodulation based on continuous models with application to speech recognition. In Proceedings of (EUROSPEECH'03), 2853-2856, Geneva, Switzerland.

[3] Dimitriadis, V., Maragos, P. and Potamianos, A. Robust AM-FM features for speech recognition. IEEE Signal Process. Letters, vol. 12, no. 9, 621-624, 2005.

[4] Francesco, G., Giorgio, B., Paolo, C. and Claudio, T. Multicomponent AM-FM representations: An asymptotically exact approach. IEEE Trans. Audio, Speech and Language Processing, vol. 15, no. 3, 823-837, 2007.

[5] Gales, M. J. F. and Young, S. J. On stochastic feature and model compensation approaches to robust speech recognition. Speech Communication, vol. 25, 29-47, 1998.

[6] Graciarena, M., Kajarekar, S., Stolcke, A. and Shriberg, E. Noise robust speaker identification for spontaneous Arabic speech. 2007. In Proceedings of IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'07), IV-245-IV-248.

[7] Grimaldi, M. and Cummins, F. Speaker identification using instantaneous frequencies. IEEE Trans. Audio, Speech and Language Processing, vol. 16, no. 6, 1097-1111, 2008.

[8] Holmes, N. J. and Sedgwick, N. C. 1986. Noise compensation for speech probabilistic models. In Proceedings of IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'86).

[9] Hung, J.-W. and Lee, L.-S. Optimization of temporal filters for constructing robust features in speech recognition. IEEE Trans. Audio, Speech and Language Processing, vol. 14, no. 3, 808-832, 2006.

[10] Islam, M. R. and Rahman, M. F. Noise robust speaker identification using PCA based genetic algorithm. International Journal of Computer Applications, vol. 4, no. 12, 27-31, 2010.

[11] Jankowski, C. R., Quatieri, T. F. and Reynolds, D. A. 1995. Measuring fine structure in speech: Application to

speaker identification. In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 325-328.

[12] Krishnamurthy, N. and Hansen, J. H. L. Babble noise: modeling, analysis, and applications. IEEE Trans. Audio, Speech and Language Processing, vol. 17, no.7, 1394-1407, 2009.

[13] Lee, C. -H. Cepstral parameter compensation for hmm recognition in noise. Speech Communication, vol. 12, 231-239, 1993.

[14] Li, G., Qiu, L. and Ng, K. L. Signal representation based on instantaneous amplitude models with application to speech synthesis. IEEE Trans. Speech and Audio Processing, vol. 8, no. 3, 353-357, 2000.

[15] Lindemann, E. and Kates, J. M. Phase relationships and amplitude envelopes in auditory perception. In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 17-20, New Paltz, New York, 1999.

[16] Maragos, P., Kaiser, J. F. and Quatieri, T. F. Energy separation in signal modulations with application to speech analysis. IEEE Trans. Signal Processing, vol. 41, no. 10, 3024-3051, 1993.

[17] Marchetto, E., Avanzini, F. and Flego, F. 2009. An automatic speaker recognition system for intelligence applications. In Proceedings of European Signal Processing Conference (EUSIPCO' 09), 1612-1616, Glasgow, Scotland.

[18] McAulay, R. J. and Quatieri, T. F. Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans. Acoustic Speech and Signal Processing, vol. 34, 744-754, 1986.

[19] Ming, J., Hazen, T. J., Glass, J. R. and Reynolds, D. A. Robust speaker recognition in noisy conditions. IEEE Trans. Audio, Speech and Language Processing, vol. 15, no.5, 1711-1723, 2007.

[20] N. I. of Standards and Technology. The NIST SRE 2008 evaluation plan (SRE-08). Technical report, 2008.

[21] Potamianos, A. and Maragos, P. A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation. Signal Processing, vol. 37, 95-120, 1994.

[22] Potamianos, A. and Maragos, P. Speech formant frequency and bandwidth tracking using multiband energy demodulation. J. Acoust. Soc. Am., vol. 99, no. 6, 3795-3806, 1996.

[23] Potamianos, A. and Maragos, P. Time-frequency distributions for automatic speech recognition. IEEE Trans. Speech and Audio Processing, vol. 9, no. 3, 196-200, 2001.

[24] Rabiner, L. R. and Shafer, R. W. Digital Signal Processing of Speech Signals. Englewood Cliffs, NJ:Prentice-Hall, 1989.

[25] Ramasubramanian, V., Vijaywargiay, D. and Kumar, V. P. 2006. Highly noise robust text-dependent speaker recognition based on hypothesized wiener filtering. In Proceedings of INTERSPEECH 2006 (ICSLP' 06), 1455-1458, Pittsburgh, Pennsylvania.

[26] Rao, A. and Kumaresan, R. On decomposing speech into modulated components. IEEE Trans. Speech Audio Processing, vol. 8, no. 3, 240-254, 2000.

[27] Reynolds, D. A. Experimental evaluation of features for robust speaker identification. IEEE Trans. Speech and Audio Processing, vol. 2, no. 4, 639-642, 1994.

[28] Reynolds, D. A. Speaker identification and verification using Gaussian mixture speaker models. Speech Communication, vol. 17, 91-108, 1995.

[29] Reynolds, D. A. An overview of automatic speaker recognition technology. 2002. In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'02), 4072-4075.

[30] Reynolds, D. A. and Rose, R. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech and Audio Processing, vol. 3, no. 1, 72-83, 1995.

[31] Tam, Y. C. and Mark, B. 2000. Optimization of sub-band weights using simulated noisy speech in multi-band speech recognition. In Proceedings of (ICSLP'00), 313-316.

[32] Varga, A. P. and Moore, R. K. 1990. Hidden Markov model decomposition of speech noise. In Proceedings of IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'90), 845-848.

[33] Wan, V. and Renals, S. 2002. Evaluation of kernel methods for speaker verification and identification. In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol. 1, 669-672.

[34] Boashash, B. 1992. Estimating and interpreting the instantaneous frequency of a signal-part 1: Fundamentals. Proc. IEEE, vol. 80, no. 4, pp. 519-538.