PITCH MAXIMA FOR ROBUST SPEAKER RECOGNITION

S Krishnakumar[*], *K R Prasanna Kumar*[†], *and N Balakrishnan* [‡]

Indian Institute of Science, Bangalore
skk@mmsl.serc.iisc.ernet.in

ABSTRACT

This paper presents a novel approach to the design of a robust speaker recognition system. A noise-free synthesised spectrum is produced from a noisy spectrum. This synthesised spectrum is used for feature extraction. From noisy speech, the pitch is extracted using a robust pitch estimation algorithm. This also helps in identifying the voiced segments of speech which are the only ones considered in the synthesis. After estimating pitch, the noisy signal is sampled in the frequency domain at pitch harmonics. From the sampled data, a reconstruction procedure is suggested in this paper in order to generate a noise-free synthesised spectrum which retains the charecteristics of the speaker but rejects the noisy contributions. We compare results with the original MFCC parameters and show that on a 100 speaker database, the MFCC parameters computed on the reconstructed spectrum consistently outperforms conventional MFCC parameters over a full range of noise levels under mismatched conditions, while maintaining comparable performance under matched conditions.

1. INTRODUCTION

Spectral domain feature extraction techniques like MFCC and PLP have proven to be quite robust to noise and are used extensively for speech and speaker recognition tasks. Although their performance is very good under clean conditions and acceptable under noisy matched conditions, it is very poor under mismatched conditions. One of the reasons for this is that the features are extracted from a spectrum corrupted by noise. On addition of noise, points of low SNR (valleys) get filled up in the spectrum, and points of high SNR (peaks) are less affected. This fact has been used in [1, 2] to extract features for robust speech recognition. Unlike speech recognition systems which make use of both voiced and unvoiced portions of speech, it is sufficient for speaker recognition systems to rely on voiced portions only [3], since they are robust to additive noise.

The motivation for the current work evolved from the following emphirical observations:

1. In clean voiced speech, the local maxima occur at pitch harmonics. Due to their higher magnitude, they are more robust to additive noise than points in their vicinity.
2. The additive noise affects the spectral valleys due to their low SNR. This is one of the primary reasons for the failure of the conventional MFCC features.
3. The width of the lobes around most local maxima in a voiced clean speech are observed to be more or less a constant.

Advances in robust pitch extraction algorithms[4, 5, 6] have made it feasible to accurately extract pitch from noisy speech. Using this pitch information, it is possible to estimate the positions of the local maxima of the clean speech spectrum from the noisy spectrum. If truncated gaussians are constructed around these positions, they would reintroduce the valleys that were filled up. This makes the MFCC parameters extracted from this reconstructed spectrum more robust. We call these parameters rMFCC.

Section 2 describes the spectrum reconstruction process, first describing the process of extracting the maxima positions (2.1), the variance (2.2) and then the reconstruction (2.3). Analysis into the shape of the magnitude spectrum for varying noise levels is presented for both the original and reconstructed spectrums. 3.1. Experiments comparing the speaker recognition performance using MFCC parameters extracted in the conventional method and using our proposed algorithm in both matched (3.3) and mismatched (3.4) environments are presented. In section 4, we present our conclusions and future directions.

2. SPECTRUM RECONSTRUCTION USING PITCH MAXIMA

2.1. Maxima extraction

The first step in the maxima extraction process is to accurately determine the pitch. We have used the algorithm proposed in [6] as our focus was on the recognition experiment. Other pitch extraction algorithms can be used too. Performance of this algorithm depends on the accuracy of the pitch extracted.

The analysis for each hamming windowed voiced speech frame, $s_i[n]$, is as follows. The pitch P_i is estimated for the ith frame under consideration. Let $H_i(f)$ be the short term Fourier spectrum of the frame from 0 to $\frac{f_s}{2}$. The sampled magnitude spectrum, $H_i^s(f)$, is constructed from $H_i[k]$, where $H_i[k]$ is obtained by computing the N point FFT of $s_i[n]$. An estimate of continuous spectrum, $\hat{H}_i(f)$, is obtained by linear interpolation of the adjacent samples of $H_i^s(f)$. An impulse train $\delta_{P_i}(f)$ is used to obtain the estimate of the corrosponding clean speech pitch maxima positions, f_m, and their corresponding points on the estimated magnitude response, $\hat{H}_i(f_m)$. To accommodate for the phase variation of these maxima, the cross correlation, $\Re_i(\tau)$, between $\hat{H}_i(f)$ and $\delta_{P_i}(f)$ is computed for all $\tau \in \tau_R$, determined by the position of the known maximum of the spectrum. The cross correlation function $\Re_i(\tau)$ is computed as

[*] Supercomputer Education and Research Centre (SERC)

[†] Dept. of Aerospace Engineering

[‡] Dept. of Aerospace Engineering and SERC

$$\begin{aligned}\Re_i(\tau) &= \int \hat{H}_i(f)\,\delta_{P_i}(f-\tau)df; \tau \in \tau_R \\ &= \sum_{k=0}^{K_i} \hat{H}_i(f-kP_i-\tau); K_i = \left[\frac{f_s}{P_i}\right]\end{aligned}$$

where

$$\tau_R \in \left[\left(\frac{k_{max} f_s}{N} - r\right) \bmod P_i, \left(\frac{k_{max} f_s}{N} + r\right) \bmod P_i\right] \quad (1)$$

k_{max} is the maximum of $H_i[k]$, and r is the frequency around $\frac{k_{max} f_s}{N}$ in which the corresponding clean speech pitch maxima is expected to lie. $\frac{k_{max} f_s}{N}$, being the global maxima of $\hat{H}_i(f)$, is expected to have highest SNR and correspondingly, it can be relied upon to be more robust to noise compared to the other points in the spectrum.

The delay, τ_i^{max} which results in $\Re_i(\tau)$ being a maximum is computed. Since the local maxima of the corresponding clean speech magnitude spectrum of the current frame occur at pitch intervals, τ_i^{max} is such that $\delta_{P_i}(f-\tau_i^{max})$ is synchronous with these maxima positions. The assumption being made here is that, taken together, the pitch maxima positions are sufficiently robust so that their exact positions can be correctly estimated.

The pitch maxima, $M_i(f_m)$, are now given as

$$\begin{aligned}M_i(f_m) &= \hat{H}_i(f_m) \\ f_m &= mP_i + \tau_i^{max}; m \in [0, K_i]\end{aligned}$$

2.2. Variance estimation

The variance of the Gaussian pulse, that best fits the frame's magnitude response, $H_i(f)$, is estimated by using the collective knowledge of lobes around the first N_B pitch maxima having maximum magnitude. Let $\{f_m, M_i\}, m \in [0, K_i]$ be rearranged to give $\{f_n, N_i\}, n \in [0, K_i]$ such that $N_i(f_j) \geq N_i(f_{j+1})$, $j \in [0, K_i - 1]$.

The segments of the sampled magnitude spectrum, $H_i^s(f)$, of width P_i Hz centred around the first N_B frequencies of $\{f_n, N_i\}$ are considered. Let $S_i^s(f;j)$ be the jth such segment of the ith frame. Then

$$S_i^s(f;j) = H_i^s(f)\texttt{rect}\left(\frac{f-f_j}{P_i}\right); j \in [0, N_B] \quad (2)$$

where $\texttt{rect}\left(\frac{f}{P_i}\right)$ is a regtangular window of unit height and width P_i, centered around the origin.

The segment values are first normalized to give

$$\bar{S}_i(f;j) = \frac{S_i(f;j)}{\int S_i(f;j)df}$$

This is done to give equal weightage to all the N_B lobes for the purpose of variance estimation.

$\bar{S}_i(f;j)$ can now be considered as a pdf. All the segments $\bar{S}_i(f;j), j \in [0, N_B]$ are shifted by $\mathbf{E}_{\bar{S}_i(f;j)}(f)$ and added to give

$$X_i(f) = \frac{1}{N_B}\sum_{j=0}^{N_B} \bar{S}_i\left(f - \mathbf{E}_{\bar{S}_i(f;j)}(f); j\right)$$

where

$$\mathbf{E}_{\bar{S}_i(f;j)}(f) = \int f\bar{S}_i(f;j)df$$

The variance of the Gaussian pulse, σ_i^2 is then calculated as

$$\sigma_i^2 = \int (f-\mu_f)^2 X_i(f)\,df$$

where $\mu_f = \int fX_i(f)\,df$. All the gaussian pulses around the extracted maxima positions 2.1 are modeled using this variance.

2.3. Reconstruction

Once the pitch maxima positions and the variance of the Gaussian pulse, σ_i^2 are estimated, the sequence of Gaussian pulses can be reconstructed. This sequence models the current frame's corrosponding clean magnitude spectrum. Let $R_i(f;j)$ be the jth reconstructed pulse in the magnitude response of the ith frame. Then,

$$\begin{aligned}R_i(f-f_j;j) &= M_i(f_j)\exp\left\{-\frac{f^2}{2\sigma_i^2}\right\}W(f) \\ \bigvee & \quad j \in [0, K_i]\end{aligned}$$

If $R_i(f)$ be the final reconstructed spectrum, then

$$R_i(f) = \sum_{j=0}^{K_i} R_i(f;j)$$

Any spectral domain based feature extraction process could be applied on this reconstructed spectrum, $R_i(f)$.

The spectrum, $\hat{H}_i(f)$, of a voiced, clean speech segment, is shown in fig 1. Also shown are the reconstructed spectrum, $R_i(f)$, and the points corresponding to $M_i(f_m)$. In fig 2, the same speech frame's magnitude response corrupted by 0dB additive Gaussian noise is shown. Also shown are the reconstructed spectrum, the original clean spectrum and the pitch points, $M_i(f_m)$.

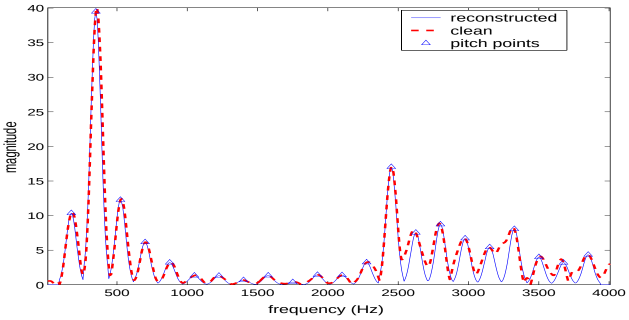

Fig. 1. Magnitude spectrum of clean speech and its reconstructed waveform.Also shown are the pitch peaks which are picked.

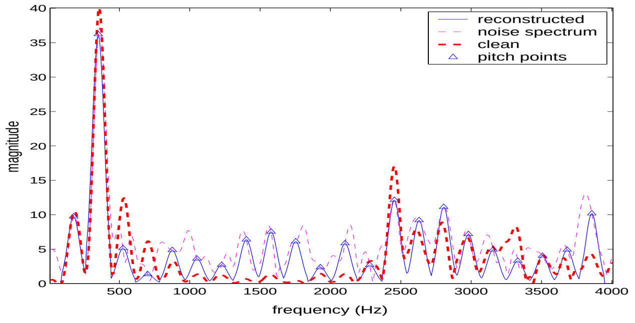

Fig. 2. Magnitude spectrum of a speech waveform, corrupted by a 0dB additive white noise. Also shown are its reconstructed waveform and the corresponding clean speech waveform. The pitch peaks picked is also shown.

3. EXPERIMENTS

3.1. Spectrum Analysis

Two sets of experiments were performed to compare the spectrum shapes of the original and reconstructed spectrums in clean and noisy environments for different noise levels. Let the original clean spectrum be indicated by cO, the original noisy spectrum by nO, the reconstructed clean spectrum by cR, and the reconstructed noisy spectrum by nR. All utterances of a single female speaker from the TIMIT database were concatenated and the magnitude spectrum of all the voiced frames were used in this experiment. nO was got by adding white gaussian noise [8] to cO to maintain the desired SNR level. cR and nR are extracted from cO and nO respectively using the proposed algorithm in 2.

In the first set of experiments, the euclidian distances, $d(cO, nO)$ and $d(cR, nR)$ are evaluated where $d(x, y)$ is the euclidian distance between x and y. The scatter plots between $d(cO, nO)$ and $d(cR, nR)$ are shown in (a)-(c) of figure 3 for noise levels of 0dB, 5dB and 10dB. From the plots, it can be seen that, on an average, $d(cR, nR)$ is around 5dB lesser than $d(cO, nO)$ for all noise levels.

In the second set of experiments, the cross-correlation coefficient, ρ, where

$$\rho(x,y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

is compared for the original waveforms, $\rho(cO, nO)$, and the reconstructed waveforms, $\rho(cR, nR)$. The scatter plots of these values for different noise levels are shown in (d)-(f) of figure 3. It is clearly seen that $\rho(cR, nR)$ is greater than $\rho(cO, nO)$ on an average. This shows that the shape of the spectrum is better retained between the reconstructed clean and noisy waveforms compared to the original clean and noisy waveforms.

3.2. Speaker Recognition Experiments

It can be observed that the proposed technique reconstructs only the voiced sections of speech. Since we are interested in the robust performance, relying on voiced sections is justified. The performance of the reconstructed spectrum for the task of speaker recognition was evaluated under matched and mismatched conditions. The MFCC coefficients were computed from the reconstructed spectrum (rMFCC) as well as the original spectrum in the conventional manner. To make the comparison rigid, one-to-one correspondance between MFCC and rMFCC features was maintained.

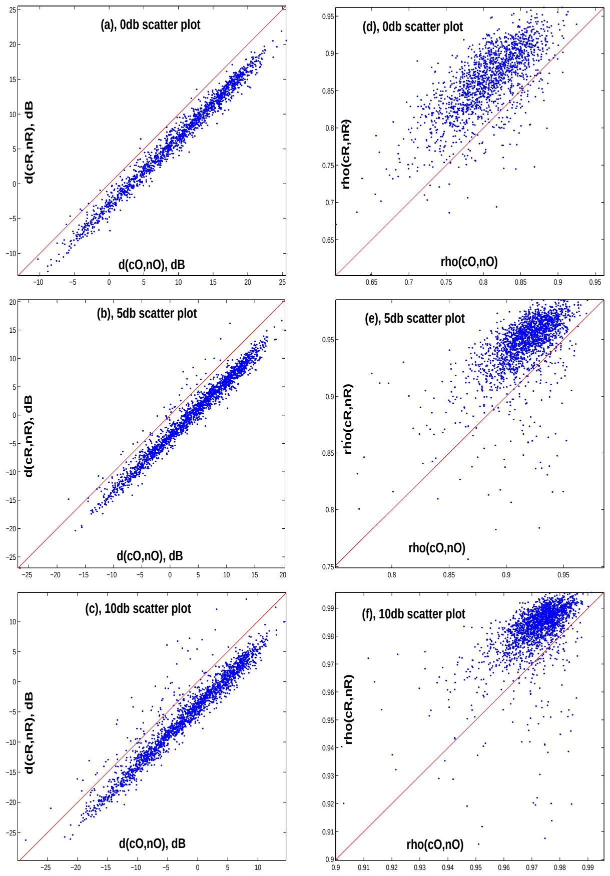

Fig. 3. (a)-(c) Scatter plots of the euclidian distance between original and reconstructed waveforms at 0dB, 5dB and 10dB SNR. (d)-(f) Scatter plots of ρ between original and reconstructed waveforms at 0dB, 5dB and 10dB SNR.

The performance was studied on speakers from the TIMIT database. The files were down sampled to 8kHz from the original 16kHz for accommodating real world situations. Considering the fact that only voiced sections were being used for training and testing, a 100 speaker subset of the 630 speakers, who had voiced segments of duration greater than 14.5 seconds were considered. All the speech files from a speaker were concatenated for this purpose. The first 11 seconds of voiced speech of individual speakers were used to model the GMMs using the MFCC and rMFCC coefficients. The last 3 seconds were used for test.

For the experiments using the reconstructed spectrum, the steps explained in the section 2 were followed. For pitch maxima extraction, r, from equ 1, was fixed at $\frac{2f_s}{N}$. For variance estimation, N_B, from equ 2, was fixed at 4. This was based on some emperical observations that the lobes around the first 4 pitch maxima (sorted in decreasing order of magnitude), appeared to be preserved under extreme noisy conditions.

The additive noise files were obtained by adding the white gaussian noise [8] to clean speech so as to maintain the segmental SNR around the desired value. The frame size of each hamming windowed speech segment was kept at 20ms with 50% overlap across frames for both the cases. Speech was not preemphasised to keep the stress on comparison of the feature space. The original and reconstructed spectrum were passed through 28 mel banks in the frequency range of 100Hz to 3400Hz. The first 17 DCT coefficients were used as features. The GMMs were used to model these MFCC coefficients using the EM algorithm[7]. In all the experiments 40 Gaussian mixtures were used for modeling.

The identification results on matched conditions were obtained by modeling the speakers under similar train and test conditions. The identification results to evaluate the robustness under mismatched conditions were obtained by training the speakers with clean speech and then testing under adverse test conditions. Five different SNR levels were considered at 0,5,10,15,25 30 dB.

3.3. Matched condition performance

The speaker recognition results for the two methods for the matched conditions are tabulated in the Table 1. The Top3 score are also tabulated. It can be seen that the rMFCC features perform comparable to MFCC features under matched conditions.

	MFCC		rMFCC	
SNR	Top1	Top3	Top1	Top3
0	68.09	85.11	72.34	85.11
5	94.00	97.00	89.00	96.00
10	92.00	100.00	94.00	98.00
15	96.00	99.00	92.00	96.00
20	93.00	98.00	91.00	96.00
25	92.00	97.00	94.00	98.00
30	95.00	97.00	95.00	98.00

Table 1. Comparison of speaker recognition performance using conventional MFCC parameters and robust MFCC parameters under matched conditions

3.4. Mismatched condition performance

This experiment compares the robustness of the two methods. The table 2 shows a comparison of the speaker recognition scores using conventional MFCC features and the rMFCC features. The Top3 scores are also tabulated. It is clear that the rMFCC features extracted from the reconstructed spectrum are more robust under mismatched conditions compared to the MFCC features.

	MFCC		rMFCC	
SNR	Top1	Top3	Top1	Top3
0	2.13	10.64	5.32	10.64
5	4.00	16.00	12.00	20.00
10	15.00	39.00	24.00	45.00
15	40.00	58.00	57.00	76.00
20	73.00	88.00	82.00	92.00
25	87.00	96.00	91.00	98.00
30	92.00	97.00	91.00	97.00

Table 2. Comparison of speaker recognition performance using conventional MFCC parameters and robust MFCC parameters under mismatched conditions. Training was done on clean speech and testing on speech corrupted by additive white noise

4. CONCLUSION

Additive noise has been shown to reduce the identification scores of the speaker recognition systems drastically under adverse test conditions. In this context, this paper has examined a robust framework for presenting the spectral information. This is achieved by presenting the spectral peaks of the voiced regions which are more robust to additive noise. The spectral peaks are computed as maxima at pitch intervals and the spectrum is reconstructed by using a robust pitch algorithm to locate the position of these maxima. Further, the rest of the spectral points are estimated by constructing Gaussians around these maxima which should serve the purpose of high reliablility under noisy conditions. The spectrum so computed has shown to perform comparable to MFCC coefficients in matched conditions. Under test conditions of additive noise, the reconstruction approach has shown to outperform the conventional MFCC features. Two major issues make the method more attractive inspite of reduced train and test data for the reasons of relying on only voiced regions. They are the tractability of pitch maxima and a scope for handling the spectral compression. In the present framework the variance of the spectral features at valleys was handled by forcing a Gaussian pulse around the maxima. The results indicate that the pitch maxima could provide useful cue for speaker recognition in robust conditions. It will be interesting to see its performance with delta coefficients. Tracking of the pitch maxima across the speech utterance could further help improve robustness of speaker recognition systems to additive noise.

5. REFERENCES

[1] Raj, B, “Reconstruction of Incomplete Spectrograms for Robust Speech Recognition”, PhD Thesis, Carnegie Mellon University, April 2000

[2] Bojana Gajic and Kuldip K. Paliwal, “Robust Feature Extraction Using Subband Spectral Centroid Histograms”, in Proc. of ICASSP ’01, vol 1, pp 85-88

[3] C. R. Jankowski Jr., T. F. Quatieri, and D. A. Reynolds, “Measuring Fine Structure in Speech: Application to Speaker Recognition”, ICASSP ’95, vol 1, pp 325-328

[4] Sun, X. “Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio”. In the Proc. of ICASSP ’02, Orlando, Florida, May 13-17, 2002

[5] P. C. Bagshaw, S. M. Hiller and M. A. Jack, “Enhanced Pitch Tracking and the Processing of f0 Contours for Computer Aided Intonation Teaching,” Eurospeech ’93, Berlin, 1993, pp. 1003-1006

[6] Y. Medan, E. Yair, and D. Chazan, “Super resolution pitch determination of speech signals,” IEEE Trans. Acoust., Speech, Signal Processing, vol. 39, pp. 40-48, Jan. 1991

[7] Douglas A. Reynolds, Richard C Rose, “Robust Text Independent Speaker Identification Using Gaussian Mixture Speaker Models”, IEEE Transactions on Speech and Audio Proc., Vol 3, No 1, January 1995

[8] J. H. L. Hansen and L. Arslan, "Robust feature- estimation and objective quality assessment for noisy speech recognition using the credit-card corpus," IEEE Trans. Speech Audio Processing," Vol. 3, pp. 169-184, May 1995