Effect of Speech Coding on Speaker Identification

Anil Kumar Vuppala
G.S.S.S.T
Indian Institute of Technology
Kharagpur, 721302, India
Email: anil.vuppala@gmail.com

K. Sreenivasa Rao
S.I.T
Indian Institute of Technology
Kharagpur, 721302, India
Email: ksrao@sit.iitkgp.ernet.in

Saswat Chakrabarti
G.S.S.S.T
Indian Institute of Technology
Kharagpur, 721302, India
Email: saswat@ece.iitkgp.ernet.in

***Abstract*—The increasing use of wireless systems is creating great deal of interest in the development of robust speech systems in wireless environment. The major degradations involved in wireless environment are: effect of varying background conditions, degradation due to speech coders and errors due to wireless channels. In this paper, we presented the effect of speech coding on text independent speaker identification (SI). Speech coders considered in this work are GSM full rate (ETSI 06.10), CELP (FS-1016), and MELP (TI 2.4kbps). The amount of distortion introduced by coding is measured using log-likelihood ratio (LLR), weighted spectral slope (WSS) and log-spectral distance (LSD). The effect of coding on SI is analyzed by building SI system using both vocal track system and excitation source features. We observed that there is a significant reduction of performance in SI system due to coding, and effect is more prominent in case of SI system build with source features. We also observed that, speaker characteristics are well preserved in case of MELP compared to CELP even though MELP coder bit rate is less than CELP.**

***Index Terms*—Speaker identification, Wireless environment, speech coders, GSM full rate (ETSI 06.10), CELP (FS-1016), and MELP (TI 2.4kbps)**

I. Introduction

Robust speech systems in wireless environment has gained a special interest in recent years in order to enable access to remote voice-activated services. In this context, three major challenges that needs to be considered are: varying background conditions, speech coding and transmission channel errors [1], [2]. Whilst the first one has already received a lot of attention, the last two deserve further investigation.

Recently, there has been increasing interest in studying the effects of speech coding on the performance of speech systems. Low-rate speech coders are very effective to maximize the utilization of expensive wireless bandwidth. This aggressive compression of speech signal will degrade the information present in speech signal (i.e., message, speaker and language information). The main goal of this paper is to analyze the effect of speech coding on text independent speaker identification (SI) system. In SI, the task is to identify the speaker from from the speech signal. Speech coders considered in this work are GSM full rate (ETSI 06.10), CELP (FS-1016), and MELP (TI 2.4kbps).

In literature the effect of coding on speaker recognition performance was analyzed in two ways. In the first case features required for speaker recognition are extracted from resynthesized speech [3]–[5], and in the second case features are extracted directly from the codec parameters [3]. Quatieri (1999) *et al* [3] studied the effect of GSM (12.2 kbps), G.729 (8 kbps), and G.723.1 (5.3 kbps) on speaker recognition. This study shows decrease in performance of recognition system with decreasing coding rate. Dunn (1999) *et al* [4] investigated the effect of speech coding on automatic speaker recognition when training and testing conditions are matched and mismatched. Matched condition (training and testing with the same coder) shows increase in the recognition performance. Dunn (2001) *et al* [5] improved speaker recognition performance under coding using score normalization.

In this work we analyzed the effect of speech coding on SI system by building the SI system using both system and source features. Linear predictive cepstral coefficients (LPCC) and Mel-frequency cepstral coefficients (MFCC) are used to represent system features. Real cepstral coefficients (RCC) extracted from linear predictive residual signal are used to represent source features. The amount of distortion introduced by coding is measured using log-likelihood ratio (LLR), weighted spectral slope (WSS) and log-spectral distance (LSD).

This paper is organized as fallows. Brief introduction to the coders considered in this work is presented in section II. In section III describes the measures used to estimate the amount of distortion introduced due to speech coding. The effect of coding on SI system is discussed in section IV. Summary and conclusions of the present work and scope for future work are mentioned in section V.

II. Speech Coders

A. Global System for Mobile (GSM 06.10) full rate coder

GSM full rate coder provides 13 kbps bit rate using regular pulse excitation and longterm prediction (RPE-LTP) techniques. GSM full rate speech encoder takes its input as a 13 bit uniform PCM signal sampled at 8 kHz. The input PCM signal processed on a frame-by-frame basis, with a frame size of 20 ms (160 samples). Bits allocation of GSM full rate coder is shown in Table I. Full Rate GSM was the first digital speech coding standard used in the GSM digital mobile phone system.

B. Codebook Excited Linear Prediction (CELP FS-1016)

CELP is based on the concept of linear predictive coding (LPC). LPC estimates the current speech sample by the linear combination of the past speech samples. In CELP, a codebook

978-1-4244-9074-5/10/$26.00 ©2010 IEEE

TABLE I
GSM FR (ETSI 06.10) BITS ALLOCATION

Parameter to be encoded	Number of bits
8 STP LAR coefficients	36
4 LTP Gains G	8
4 LTP Delay D	28
4 RPE Grid positions	8
4 RPE Grid block maxi-ma	24
Amplitudes	156
Total no.of bits per frame	260
Transmission bit rare	13 kbps

of different excitation signals is maintained at the encoder and decoder. The encoder finds the most suitable excitation signal and sends its index to the decoder, which then uses it to reproduce the signal. Hence the name codebook excited is given to this coder. CELP FS-1016 operates at a bit rate of 4.8 kbps. CELP is widely used speech coding algorithm, and one of practical application of it is in selective mode vocoder (SMV) for CDMA. Bits allocation for CELP FS-1016 coder is shown in Table II

TABLE II
CELP (FS-1016) BITS ALLOCATION

Parameter to be encoded	Number of bits
LPC coefficients (10 LSP's)	34
Pitch prediction	48
Code book	36
Gains	20
Synchronization	1
+91 9734532849. FEC	4
Frame expansion	1
Total no.of bits per frame	144
Transmission bit rare	4.8 kbps

C. Mixed Excited Linear Prediction (MELP TI 2.4 kbps.)

MELP utilizes more sophisticated speech production model, with the additional parameters to capture the underlying signal dynamics with the improved accuracy. The basic idea is that the mixed excitation signal is generated by combining filtered periodic pulse sequence with the filtered noise sequence and this mixed excitation signal is given input to the synthesis filter. MELP (TI 2.4 kbps) operates at a bit rate of 2.4 kbps. MELP is used in military, satellite, and secure voice applications. Bits allocation of MELP coder is shown in Table III. Finally the comparison of speech coders considered in this study in terms of complexity, Mean opinion score (MOS) and frame size is shown in Table IV.

III. DEGRADATION MEASURES

Degradation introduced by speech coding is measured using log-likelihood ratio, weighted spectral slope measure and log-spectral distance [6]. They are defined as follows.

TABLE III
MELP (TI 2.4 KBPS) BITS ALLOCATION

Parameter to be encoded	Number of bits
LPC coefficients (10 LSP's)	34
Gain (2 per frame)	8
Pitch and overall Voicing	7
Band pass voicing	5-1
Aperiodic flag	1
Total no.of bits per frame	54
Transmission bit rare	2.4 kbps

TABLE IV
COMPARISON OF CODERS

Algorithm	Bit-rate (kbps)	MOS	Complexity (MIPS)	Frame size (ms)
PCM	64	4.3	0.01	0
GSM FR	13	3.5-3.9	5-6	20
CELP	4.8	3.2	16	30
MELP	2.4	3.2	40	22.5

A. LLR–Log-Likelihood Ratio

$$LLR = log_{10}[\frac{a_x R_x a_x^T}{a_y R_y a_y^T}] \tag{1}$$

where
a_x and a_y are LPC vectors of the original and coded (degraded) speech. a_x^T and a_y^T are Transpose of LPC vectors of the original and coded (degraded) speech. R_x and R_y are autocorrelation matrices of the original and coded (degraded) speech

B. WSS–Weighted Spectral Slope Measure

The WSS measure proposed by Klatt is based on an auditory model in which 36 overlapping filters of progressively larger bandwidth are used to estimate the smoothed short-time speech spectrum

$$WSS = K_{spl}(k - \hat{k}) + \sum_{k=1}^{36} W_a(k)[S(k) - \hat{S}(k)]^2 \tag{2}$$

where k, $\hat{k}$ are related to overall sound pressure level of the original and coded (degraded) utterances. K_{spl} is a parameter which can be varied to increase overall performance. $W_a(k)$ is the weight of each band. $S(k)$ $\hat{S}(k)$ are the slopes in each critical band k for the original and coded (degraded) speech.

C. LSD–Log-Spectral Distance

$$LSD = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\log_{10}\frac{P(\omega)}{\hat{P}(\omega)}\right]^2 d\omega}. \tag{3}$$

where $P(\omega)$ and $\hat{P}(\omega)$ are power spectrum of clean and coded (degraded) speech

Above quality measures are calculated for 50 female speakers and 50 male speakers from TIMIT database [7], and average measures for both female and male speakers are shown in Tables V and VI respectively. In Tables V and VI column-1 indicates different coders considered in this study; Columns 2–4 indicates the average LLR, WSS and LSD values for different coders. Results indicating that CELP coder introduces more degradation among three coders considered. Compared to male speakers speech utterances, female speakers speech utterances are more degraded due to speech coding (see Tables V and VI.

TABLE V
QUALITY MEASURES COMPARISON OF CODED FEMALE SPEAKER UTTERANCES

Algorithm	LLR	WSS	LSD
GSM FR (ETSI 06.10)	0.22	18.31	0.94
CELP (FS1016)	0.58	52.61	1.45
MELP (TI 2.4kbps)	0.28	37.78	1.16

TABLE VI
QUALITY MEASURES COMPARISON OF CODED MALE SPEAKER UTTERANCES

Algorithm	LLR	WSS	LSD
GSM FR (ETSI 06.10)	0.22	15.4	0.89
CELP (FS1016)	0.55	43.6	1.4
MELP (TI 2.4kbps)	0.27	32.4	1.07

IV. EFFECT OF SPEECH CODING ON TEXT INDEPENDENT SPEAKER IDENTIFICATION

A. Speaker identification Experimental setup

TIMIT database [7] is used for building the speaker identification system. Randomly chosen fifty male and fifty female speakers are considered in the present study. Database contains 10 utterances from each speaker, in that 8 utterances are used for building the models and 2 utterances are used to test the models. Features extracted from every 20 ms of utterance with 10 ms frame shift are used for training and testing the acoustic models. Gaussian mixture models (GMM) with 16 mixtures are used to build the speaker models. Speaker identification performance under GSM full rate, CELP and MELP coders is measured by training the GMM models with clean speech and matched (corresponding coded) speech.

Both system and source features [8] are used to build acoustic models. Linear predictive cepstral coefficients (LPCC) and Mel-Frequency Cepstral Coefficients (MFCC) are used to represent system features, and real cepstral coefficients (RCC) are used to represent source features. For calculation of LPCCs, linear prediction (LP) analysis of order 26 is performed on the speech signal using frame size of 20 ms with frame shift of 10 ms. LP coefficients are converted into cepstral domain. We have used a 13-dimensional LPCC vector as feature vector of each frame. For calculation of MFCCs, the short term Fourier transform (STFT) analysis is performed on the speech signal using frame size of 20 ms with frame shift of 10 ms. For each frame the STFT magnitude spectrum is computed and is further processed by using the 24 triangular shaped mel-filter banks to find out the filter bank energies. Then discrete cosine transform (DCT) is taken on the spectral energies to obtain MFCCs. We have used a 13-dimensional MFCC vector (excluding c_0) as feature vector of each frame. For calculation of RCCs, LP analysis of order 8 is performed on the speech signal using frame size of 20 ms with frame shift of 10 ms. For each frame LP residual is calculated by using inverse filtering. Cepstral analysis is performed on residual to obtain RCCs. We have used a 16-dimensional RCC vector as feature vector of each frame.

B. Results and discussion

Speaker recognition performance using MFCC, LPCC and residual RCC are shown in Tables VII, VIII and IX respectively. In Tables VII, VIII and IX column-1 indicates the coding technique used for testing; column-2 indicates the performance of speaker identification models trained using PCM (clean) speech and tested by different coded speech; and column-3 indicates the performance of speaker identification models trained and tested with the corresponding coders. Results indicating that performance of the SI system is decreasing as coding rate decreases. When training and testing conditions are matched then there is an increase in the SI performance.

From the results it is evident that source features are more degraded compared to system features due to coding. We also observe that, even under matched training and testing condition also SI performance is not much improved using source features (see Tables VII, VIII and IX). So the speaker discriminative information present in source features is more degraded due to coding compared to system features. Reason for this is while coding, excitation signal is represented with less number of bits. Results also indicating that speaker specific information is well preserved in case of MELP compared to CELP coder. From the results it is evident that coding has significant effect on speaker recognition performance, so there is a need to find robust features and modeling techniques to improve the performance.

TABLE VII
SPEAKER RECOGNITION PERFORMANCE USING MFCC FEATURES UNDER DIFFERENT SPEECH CODERS

	Recognition performance (%)	
Coders	PCM training	Matched training
PCM (Clean)	88.78	-
GSM	70	82.63
CELP	51.04	77.08
MELP	75.79	86.32

V. SUMMARY AND CONCLUSIONS

In this paper, we have studied the effect of coding on text independent speaker identification. Degradation introduced

TABLE VIII
SPEAKER RECOGNITION PERFORMANCE USING LPCC FEATURES UNDER DIFFERENT SPEECH CODERS

Coders	Recognition performance (%)	
	PCM training	Matched training
PCM (Clean)	89.3878	-
GSM	47.3684	84.21
CELP	43.75	75
MELP	50	80

TABLE IX
SPEAKER RECOGNITION PERFORMANCE USING RCC FEATURES FROM LP RESIDUAL UNDER DIFFERENT SPEECH CODERS

Coders	Recognition performance (%)	
	PCM training	Matched training
PCM (Clean)	85.8586	-
GSM	14.6465	53.5354
CELP	10.6061	45.4545
MELP	13.6364	63.1313

due to speech coding is measured using log-likelihood ratio, weighted spectral slope and log-spectral distance, and it is observed that CELP coded speech is more degraded among the coders considered in this study. Degradation measures also indicating that female speakers speech is more degraded compared to male speakers. In this work we built the speaker identification system using both system features (MFCC and LPCC) and source features (RCC) to analyze the coding effect. From the results we observed that source features are more degraded due to coding compared to system features, reason for this is due to the allocation of less number of bits for excitation signal while coding the speech signal. Interesting observation in this work is MELP coders are giving less degradation and high speaker identification performance compared to CELP coder, even though MELP works at less bit rate compared to CELP. We also observed that there is a close similarity between degradation measures and speaker identification performance. Finally, from this work it is evident that coding has significant effect on speaker identification performance, so there is a need to find robust features under coding or modeling techniques to improve the speaker identification performance.

REFERENCES

[1] Tan and Zheng-Huag, "Automatic speech recognition on mobile devices and over communication networks", *Springer London*, 2008.

[2] J. Vicente-Pena and A. Gallardo-Antoln and C. Pelaez-Moreno and F. Daz-de-Mara, "Band-pass filtering of the time sequences of spectral parameters for robust wireless speech recognition", *Speech Communication*, vol. 48, pp. 1379–1398, 2006.

[3] F. Quatieri and E. Singer and R.B. Dunn and D.A. Reynolds and J.P. Campbell, "Speaker and language recognition using speech codec parameters", *Proceedings of Eurospeech*, pp. 787–790, 1999.

[4] Dunn R.B. and Quatieri T.F. and Reynolds D.A. and Campbell J.P., "Speaker recognition from coded speech in matched and mismatched condition", *Speaker Recognition Workshop01, Crete, Greece*, pp. 115–120, 1999.

[5] Dunn R.B. and Quatieri T.F. and Reynolds D.A. and Campbell J.P., "Speaker recognition from coded speech and the effects of score normalization", *IEEE Conference Record of the Thirty-Fifth Asilomar Conference on Signals, Systems and Computers*, pp. 1562–1567, 2001.

[6] John H. L. Hansen and Bryan L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms", *Proceedings of the international conference on speech and language processing*, pp. 2819–2822, 1998.

[7] J. S. Garofolo et al., TIMIT Acoustic-Phonetic Continuous Speech Corpus Linguistic Data Consortium. Philadelphia, PA, 1993.

[8] Joseph W. Picone, "Signal modeling techniques in speech recognition", *Proceedings of the IEEE*, vol. 81, pp. 1215–1247, 1993.

[9] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Trans. Speech Audio Process*, vol. 3, pp. 72–83, 1995.