Improving performance of HMM-based ASR for GSM-EFR speech coding

Lallouani Bouchakour, Mohamed Debyeche

الملخّــص

التّعرّف الآلي للكلام عن طريق الهاتف النّقّال يتأثّر بعدّة عوامل؛ منها : الترميز (effet du codec) وقناة الاتّصال، وحتّى ضجيج المحيط الذي نتكلّم فيه، إلخ. لتحسين أداء التّعرّف الآليّ على الكلام، قمنا في دراستنا هذه باستعمال مقاربتين؛ أولاهما تعتمد العوامل (les parametes MFCC) والأخرى توظّف المعاملات (LSF) ومشتقّاتها : PCC, PCEP, LPC، التي استخرجت مباشرة من الحزم البيتيّة (train des bits). للتّعرّف الآليّ على الكلام استُعملت نماذج «ماركوف» (Markov) إضافة إلى قاعدة البيانات «ARADIGT» المكوّنة من عشرة حروف، سجّلت من كلام 60 متكلّما جزائريّا، من مختلف الأعمار؛ 31 منهم ذكور و29 إناث. في النّتيجة توصّلنا إلى أنّ نتائج المقاربة الثّانية كانت أحسن مقارنة بالأولى.

الكلمات المفاتيح : التّعرف الآلي على الكلام، الهاتف النقال، MFCC, LSF، الحزم البيتية، نماذج ماركوف، ARADIGT.

Improving performance of HMM-based ASR for GSM-EFR speech coding

Lallouani Bouchakour,[1,2] Mohamed Debyeche[2]

[1]Centre de Recherche Scientifique et Technique pour le Développement de la Langue Arabe(CRSTDLA), Algiers
[2]Faculty of Electronics and Computer Sciences, LCPTS, USTHB, Algiers
lbouchakour@usthb.dz, mdebyeche@gmail.com

Abstract

The Global System for Mobile (GSM) environment includes three main problems for Automatic Speech Recognition (ASR) systems: noisy scenarios, source coding distortion and transmission errors. The second, source coding distortion must be explicitly addressed. In this paper, we investigate different features extractions techniques for GSM EFR (Enhanced Full Rate) coding with the aim to improve the performance of ASR in the GSM domain. Specifically, we suggest extracting the recognition feature vectors directly from the encoded speech bit-stream instead of decoding it and subsequently extracting the feature vectors. The speaker-independent recognition experiment was based on the Continuous Hidden Model Markov (CHMM). The performance of the proposed speech recognition technique was assessed using the ARADIGT transcoding with its 8 kHz downsampled version. Different experiments were carried out in order to explore feature calculation directly from the GSM EFR encoded parameters and to measure the degradation introduced by different aspects of the coder. The ARADIGIT database consists of 60 speakers (31 male speakers and 29 female speakers) pronouncing the ten Arabic digits, was built in order to conduct the necessary experiments. As a result, the proposed methods achieved higher performances in recognition accuracy, compared with the conventional methods employing Mel-Frequency Cepstral Coefficients MFCC. This paper presents two configurations used for extracting feature parameters for speech recognition over mobile communication; the decoded speech-based technique and the bit-stream-based technique.

Key words: speech coding, GSM, EFR, CHMM, ASR, ARADIGT, MFCC, bit-stream

1. Introduction

Fast development of mobile communication networks during the last decade opened up a new domain of expansion for speech analysis technologies. In this respect, Automatic Speech Recognition has been, and still is, a very active research topic aiming at providing easy and natural user access ways to network communication services, therefore encouraging the development of man–machine communication. This natural way of interaction has many applications because of the fast de-

Parameter	1st and 3rd subframe	2nd and 4rd subframe.	Total per frame 20ms
2 LSP sets			38
ACB index (lag)	9	6	30
ACB gain	4	4	16
FCB pulses	35	35	140
FCB gain	5	5	20
Total			244 bits

Table 1. Bit-allocation parameter of the GSM EFR codec.

velopment of different hardware and software technologies. The most relevant ones are: easy and direct access to information systems, assistance to handicapped persons and oral command systems, etc. Mobile communication networks GSM typically change within several inherent constraints (ambient noise, high compression ratio and channel noise) that inevitably tend to degrade performances of speech recognition systems. This paper addresses the effect of GSM EFR speech coding on speech recognition performance. The implemented speech recognition system is based on the CHMM probabilistic classification method. Our experiments are done on the ARADIGIT corpus. This paper is organized as follows: after this introductive section 1; the GSMEFR codec is presented in section 2; the developed speech recognition system is presented in section 3; the conducted experiments and the obtained results are given in section 4. Finally, conclusions and future work are summarized in Section 5.

2. GSM Speech Coders

Three norm coders are standardized for use in GSM communication networks. They are referred to as the Full Rate, Half Rate and Enhanced Full Rate GSM coders. Their corresponding European telecommunications standards are the GSM 06.10, GSM 06.20 and GSM 06.60, respectively. These coders work on a 13 bits uniform PCM speech input signal, sampled at 8 kHz [3]. The GSM EFR speech codec is based on the ACELP algorithm (Algebraic Code Excited Linear Prediction) [2]. The speech coding (source coding) bit-rate is 12.2 Kbit/s. For channel coding (error protection) 10.6kbit/s bit-rate is used, resulting in 22.8 Kbit/s channel bit-rate. Bit allocation of the GSM EFR codec is shown in Table I [1-3].

The EFR codec operates on 20 ms speech frames which are divided into four subframes at 5 ms. In the encoder, the speech signal is analyzed and the parameters of the ACELP speech synthesis model are extracted. The set of linear prediction filter coefficients are calculated for each frame. The indices for the adaptive (ACB) and fixed codebooks (FCB) as well as their gains are extracted for each subframe [2].

The function of the decoder consists of decoding the transmitted parameters (LSF parameters, adaptive codebook vector, adaptive codebook gain, fixed codebook vector, fixed code book gain) and performing synthesis to obtain the reconstructed speech.

3. Speech Recognition System

Nowadays, ASR systems are primarily based on the principles of statistical pattern recognition, in particular the use of Hidden Markov Models (HMMs).

The HMM is a powerful statistical method for characterizing the observed data samples of a discrete-time series.

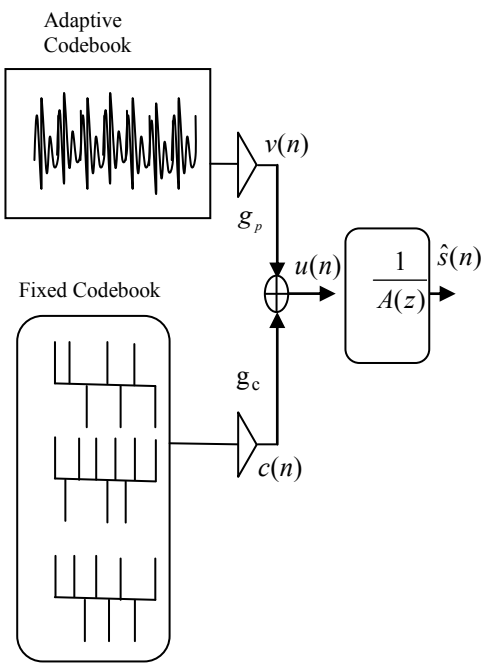

Figure 1: Simplified block diagram of the ACELP synthesis.

The underlying assumptions for applying HMMs to ASR are that the speech signal can be conveniently characterized as a parametric random process and that the parameters of the process can be precisely estimated. The architecture of a typical ASR system, depicted in figure .2, shows a sequential structure of ASR including such components as speech signal capturing, front-end feature extraction and back-end recognition decoding. Feature vectors are first extracted from the captured speech signal and then delivered to the ASR decoder. The decoder searches for the most likely word sequence that matches the feature vectors on the basis of the acoustic model. The output word sequence is then forwarded to a specific application [4].

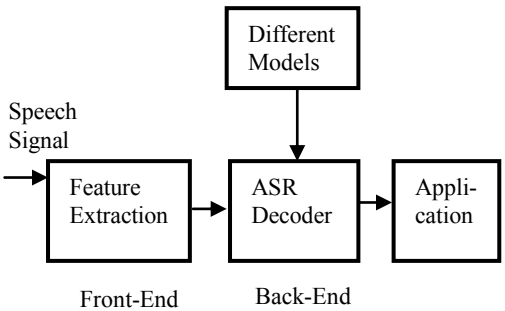

Figure 2: Architecture of an ASR system.

3.1. Feature extraction

The major reason for this function is that the parameterization of speech; for speech coding is different from that speech recognition. For example, speech coding is mainly based on a speech production model, which represents the spectral envelope of speech signals using linear predictive coding (LPC) coefficients, while feature representations used for speech recognition like, for example, Mel-frequency spectral coefficients (MFCC), are usually extracted on the basis of human perception [8].

3.2. Decoded speech-based technique

The MFCC (Mel-Frequency Spectral Coefficient) feature is a representation of the short term power spectrum of a sound.

The frequency bands in MFCC are equally spaced on the *mel scale* which closely approximates the human auditory system

response. The Mel scale can be calculated by Eq. (2) [8].

$$Mel\ (f) = 2595 \times \log_{10}(1 + \frac{f}{700}) \quad (2)$$

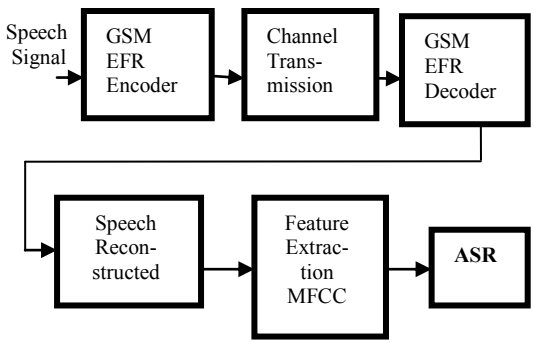

Figure 3: Architecture of a mobile network communication speech recognition system.

Mel-Frequency spectral coefficients (MFCCs) are commonly computed from FFT power coefficients filtered by a tri-angular band pass filter bank, where Aj the output of the j-th filter bank and N is the number of samples in a basic unit [8].

$$C_n = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} A_j \cos\left(\frac{\pi n}{N}(j - 0.5)\right) \quad (3)$$

3.3. Bitstream-based technique

This allows obtaining speech recognition parameters directly from the bit-stream transmitted to the receiver over digital mobile networks. This technique is based on the decoded speech to avoid the stage of reconstructing speech from the coded speech parameters. In this scenario, we are transforming the speech coding parameters to speech recognition parameters as entered to the ASR systems [9-10].

3.3.1. Pseudo-Cepstral Coefficients (PCC)

The PCC parameters are computed directly from the LSF parameters.

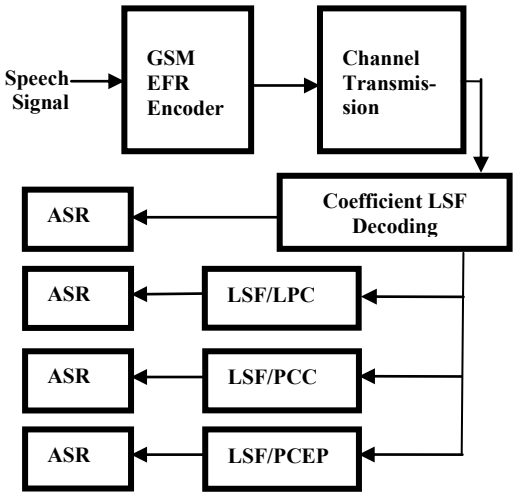

Figure 4: Feature transforms from LSFs to ASR feature parameters.

Mathematical manipulations and approx-imations allow it to be expressed in terms of the LSFs [10]. The n-th PCC is given by the equation:

$$C_n^{PCC} = \frac{(1 + (-1)^n)}{2n} + \frac{\sum_{i=1}^{p} \cos(n w_i)}{n} \quad (4)$$

Where wi is the i[th] LSF parameters, N the order of th PCC parameters, p the order of th LSF parameters.

3.3.2. Pseudo-Cepstrum Coefficients (PCEP)

Using the mathematical expression of the PCC parameters, it is somewhat trivial to obtain the PCEP. They are derived from the PCC parameters by eliminating the term:

Nb. Of Gauss.	1	2	4	8	10	12
MFCC Org. Sp.	83	92,67	91,33	92	90,83	92,5
MFCC Trans. Sp.	66	77,5	77,83	75,83	76,67	74,67

Nb. of Gauss. : Number of Gaussian
Org. So. : Original Speech
Trans. Sp. : Transcoded Speech

Table 2. Recognition results of speech original and speech GSM EFR transcoded with different number of Gaussians.

$$\frac{(1+(-1)^n)}{2n} \quad (5)$$

The n-th PCEP parameters expression is given by:

$$c_n^{PCEP} = \frac{\sum_{i=1}^{P} \cos(nw_i)}{n} \quad (6)$$

4. Experimental Results

4.1. Databases

The database used in this work is ARADIGIT database. It consists of the 10 Arabic digits spoken by 60 speakers of both sexes. This database was recorded by Algerian speakers aged between 18 and 50 years in a quiet environment. With an ambient noise level below 35 dB, in ".Wav" file format, with a sampling frequency of 8 KHz.

This database is divided into two corpuses, in speaker-independent models:

4.1.1. Train corpus

Consisting of 1200 speech files pronounced by 40 speakers including the two genders, where, each speaker repeats the 10 Arabic digits three times.

4.1.2. Test corpus

Consisting of 600 files pronounced by 20 speakers including the two sexes, where each speaker repeats the 10 Arabic digits three times.

The results of Table 2 show that recognition rate of the transcoded speech decreases compared to the rate of original speech. The optimal system performance corresponds to the 4th Gaussian number.

This can be explained by the degradation of signal quality caused by the GSM EFR coder. This degradation is created by two types of quantification by two codebooks excitation fixed and adaptive, and the quantification of spectral parameters LSF, because, it means reconstructed is the result of convolution the excitation signal (two codebooks fixed and adaptive) and vocal tract parameters. This convolution deforms the spectral envelope.

4.2. Bit-stream recognition technique

The ASR feature analysis can be extracted directly from the bit-stream parameters, as produced by the channel decoder in a mobile cellular communications network.

Nb. of Gauss.	1	2	4	8	10	12
LPC	66,17	72,33	68,5	73,17	70,83	70,33
LSF	71,17	79,17	82,5	82,33	81,83	82,5
PCC	80,33	86,33	87,83	89,17	88,5	89,67
PCEP	81,83	87,33	88,17	87,5	87,67	88,5

Nb. of Gauss.: Number of Gaussian

Table 3. Bit-stream recognition parameters with different number of Gaussians.

The results of Table 3 show that the recognition rate of the bit-stream-based technique increases compared to the decoded speech-based technique.

The bit-stream-based technique avoids the stage of reconstructing speech from the coded speech parameters. The speech reconstruction is the results of convolution of the spectral envelope parameters with the quantized excitation signal. This convolution creates the source of the performance degradation in ASR systems. The transformation of speech coding parameters to speech recognition parameters motivated by the fact that ASR feature parameters are based on the speech spectral envelope and not on the excitation. The features LSF, PCC and PCEP Bit-stream techniques are required to improve ASR performance in comparing with the MFCC features of the decoded speech-based techniques.

5. Conclusion

In this paper, we investigate the influence of GSMEFR speech coding on a text-independent speech recognition system based on CHMM classifier. Our objective is to study various speech features that must be addressed to facilitate robust automatic speech recognition over mobile communication networks, with the aim of minimizing the impact of degradation performance ASR introduced by speech coder. Obtaining ASR features directly from the bit-stream technique of standardized speech coder was originally developed as a new paradigm for feature extraction over mobile communications networks.

The bit-stream parameters LSF, PCC and PCEP allow improving ASR performance in comparison with the MFCC parameters of the decoded speech technique. The bit-stream technique seems to be promising. Therefore further developments based on this technique are being conducted in our team.

To improve system performance Speech Recognition in a GSM communication network must include ways to reduce noise coding and methods of speech enhancement to enter of the recognition system.

6. References

[1] Honkanen, T., Vainoi, J., Jarvinen, Haavisto, P., Salami, R., Laflamme, C. and Adoul, J-P., "Enhanced Full Rate speech code for is-136 digital cellular system," IEEE. vol.2. pp.731 -734. 1997.

[2] Jarvinen, K., Vainio, J., Kapanen, P., Honkanen, T., Haavisto. P., Salami, R., Lajlamme, C. and Adoul J-P. "GSM Enhanced Full Rate speech codec," IEEE, pp771 - 774, 1997.

[3] Salami, R., Laflamme, C., Bessette, B. and Adoul, J-P., "Description of GSM

Enhanced Full Rate speech codec," IEEE, pp. 725- 729. 1997.

[4] Antonio, M., Peinado, J. and Segura, C., "Speech recognition over digital channels John Wiley & Sons Ltd, vol. pp 7-29, 2006.

[5] Gernot A. Fink "Markov Models for Pattern Recognition," Springer. vol. pp. 61-92. 2008.

[6] Zheng-Hua, T. and Lindberg, B., "Automatic speech recognition on mobile devices and over communication networks," Springer, vol. pp 41-58, 2008.

[7] Sadaoki, F., "Digital speech processing, synthesis and recognition," Second Edition, pp 243-328. 2001.

[8] Holmes, J. and Holmes, W., "Speech synthesis and recognition", Taylor & Francis e-Library, Second Edition, vol. pp 161-164, 2003.

[9] Fabregas, V., de Alencar, S. and Alcaim, A., "Transformations of LPC and LSF parameters to speech recognition features," Springer, vol. pp. 522-528, 2005.

[10] Hong, K. K., Seung, H. C. and Hwang S. L. "On Approximating Line Spectral Frequencies to LPC Cepstral Coefficients," IEEE, vol.8, no.2, 2000.

[11] Fabregas, V., de Alencar, S. and Alcaim, A., "On the Performance of ITU-T G.723.1 and AMR-NB Codecs for Large Vocabulary Distributed Speech Recognition in Brazilian Portuguese," IEEE, pp 693-697, 2009.