IJIREEICE

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 4, Issue 5, May 2016

Speaker and Language Recognition

Mrs. Minaj S. Shikalgar[1], **Mr. N.B.Sambre**[2]

Department of Electronics & Telecomm., KIT'S College of Engineering, Kolhapur[1,2]

Abstract: This paper presents the speaker recognition and language recognition system. Speaker recognition is done using three coders and the performance or accuraccy for each coder is calculated. Speaker recognition is the process to identify a person using characteristics extracted from their voices. While language recognition is the process to identify the language of speaker. In this paper, speaker recognition is done using three different coders at different bit rate. Here GSM at 12.2 kb/s, G.729 at 8 kb/s, and G.723.1 at 5.3 kb/s speech coders are used[3]. The task of language recognition is done using phonetic approach[10].

Index Terms: GSM at 12.2 kb/s, G.729 at 8 kb/s, and G.723.1 at 5.3 kb/s, phonetic approach.

I.INTRODUCTION

The task of speaker identification is to determine the identity of a speaker by machine. To recognize voice, the voices must be familiar in case of human beings as well as machines. The process of "getting to know" speaker is referred to as training and consists of collecting data from utterances of people to be identified. The second component of speaker identification is testing; namely the task of comparing an unidentified utterance to the training data and making the identification. The speaker of a test utterance is referred to as the target speaker. The problem of speaker identification can be solved by using different speech coders[3]. Language can be recognized using phonetic approach[10]. The phonotactic language recognition systems always use Phonetic Recognizer (PR) as a first block which transforms speech into a sequence of phonetic labels (this is actually the first block of any speech recognition system), and operate from this point only on those phonetic labels. Systems can use a single PR or many different PRs from different languages (Parallel PR, or PPR) for better performance.

II. SPEAKER RECOGNITION

Input:
Input to this coder is a speech file in wave format and sampled at 8Khz sampling rate.

A. coders

The coders operate with a digital signal obtained by first performing filtering of the analogue input, then sampling at 8000 Hz and then convert to 16-bit linear PCM for the input to the encoder. The output of the decoder is converted back to analogue.

The G.729 coder is a fixed point codec at 8 kb/s which is based on the Code-Excited Linear-Prediction (CELP) coding model[2]. The coder operates on speech frames of 10 ms corresponding to 80 samples at a sampling rate of 8000 samples per second. For every 10 ms frame, the speech signal is analyzed to extract the parameters of the CELP model (linear-prediction filter coefficients, adaptive and fixed-codebook indices and gains). These parameters are encoded and transmitted.

The GSM Coder is ETSI Pan European standard fixed point codec which is classified as Half Rate at 5.6 kb/sand full rate at 12.2 kb/s[1]. This coder belongs to the class of Regular Pulse Excitation - Long Term Prediction -linear predictive (RPE-LTP) coders. In the encoder part, a frame of 160 speech samples is encoded as a block of 260 bits, leading to a bit rate of 13 kbps. The decoder maps the encoded blocks of 260 bits to output blocks of 160 reconstructed speech samples.

The G.723.1 coder at 5.3 kb/s is based on the principles of linear prediction analysis-by-synthesis coding and attempts to minimize a perceptually weighted error signal. The encoder operates on blocks (frames) of 240 samples each[4]. That is equal to 30 msec at an 8 kHz sampling rate. First each block is high pass filtered to remove the DC component and then divided into four subframes of 60 samples each. For every two subframes (120 samples), the open loop pitch period, is computed using the weighted speech signal. This pitch estimation is performed on blocks of 120 samples. The pitch period is searched in the range from 18 to 142 samples. From this point the speech is processed on a 60 samples per subframe basis. Using the estimated pitch period computed previously, a harmonic noise shaping filter is constructed. The combination of the LPC synthesis filter, the formant perceptual weighting filter, and the harmonic noise shaping filter is used to create an impulse response. The impulse response is then used for further computations. Using the pitch period estimation, *LOL*, and the impulse response, a closed loop pitch predictor is computed. A fifth order pitch predictor is used.

The pitch period is computed as a small differential value around the open loop pitch estimate. The contribution of the pitch predictor is then subtracted from the initial target vector. Both the pitch period and the differential value are transmitted to the decoder. Finally the non-periodic component of the excitation is approximated. For the high bit rate, Multi-pulse Maximum Likelihood Quantization (MP-MLQ) excitation is used, and for the low bit rate, an algebraic-code-excitation (ACELP) is used. The G.723.1 codec is the floating point1 CELP-based ITU-T multi-media standard codec at 5.3 kb/s.

DOI 10.17148/IJIREEICE.2016.4511

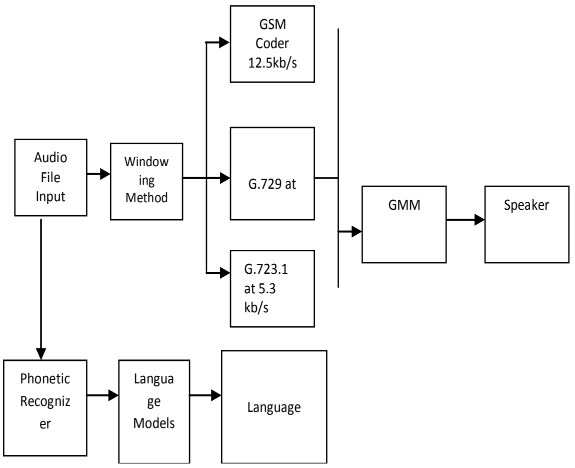

FIGURE 1 SPEAKER & LANGUAGE RECOGNITION SYSTEM

Under three different conditions the coders are tested against a baseline condition in which no coding is performed for training and testing [3].

- Condition A: Fully matched
- Condition B : Partially mismatched
- Condition C : Fully mismatched

In speaker recognition the experimentation is done with 10 different subjects. Among which 8 are adults. and 2 are kids. Again in that 8 adults subjects 4 are male and 4 are female. Similarly, in 4 kids 2 are boys and 2 are girls.The individual recording was recorded. The performance of the system totally depends on the bit rate of that coder which is shown in the following figure

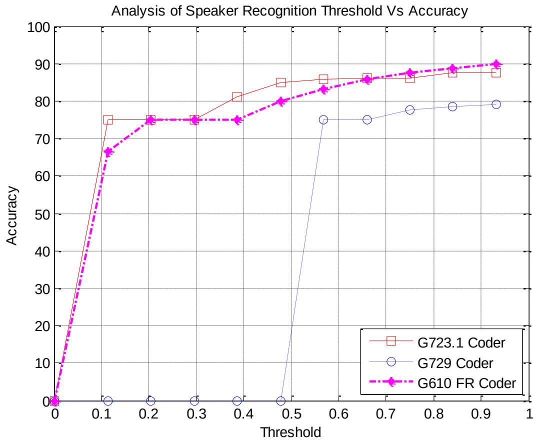

FIGURE 2 COMPARATIVE ANALYSIS OF CODERS .

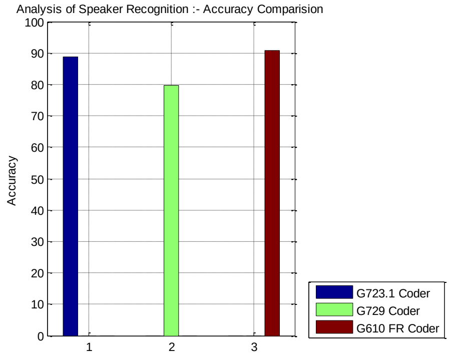

FIGURE 3 ACCURACCY FOR EACH CODER

III. LANGUAGE RECOGNITION

The phonotactic language recognition systems always use Phonetic Recognizer (PR) as a first block which transforms speech into a sequence of phonetic labels (this is actually the first block of any speech recognition system), and operate from this point only on those phonetic labels. Systems can use a single PR or many different PRs from different languages (Parallel PR, or PPR) for better performance[10]. The speakers speech is recorded in three different languages.

A PPRLM system is the combination of several PRLM systems in parallel. The building of a PRLM system starts by training a Universal Background Model (UBM) intended to represent the generality of all languages from the phonetic sequences obtained from utterances in many languages. Models for each language (*LMi*) are trained using many phonetic sequences obtained from utterances in that particular language. In most cases, these language models are adapted from the UBM to increase robustness in parameter estimation. Once the statistical language models are trained, the procedure is to verify a test utterance against a language model *LMi* using PRLM [10].

The target languages selected are English, Hindi and Marathi. The accuracy of the system is 61.1111.

IV. CONCLUSION

The performance of the system totally depends on the bit rate of that coder. In the part of language recognition, the base system of this work was a PPRLM, which has PRs and the accuracy of language recognition system which is obtained is 61.1111.

REFERENCES

[1] J.M. Huerta and R.M. Stern, \Speech recognition from GSM coder parameters," Proc. 5th Int. Conf. on Spoken Language Processing, Vol 4, pp 1463-1466, 1998.

[2] ITU-T Recommendation G.729, \Coding of speech at 8 kb/s using conjugate-structure algebraic-code-excited linear pre- diction," June 1995.

[3] M.A. Zissman, \Predicting, Diagnosing, and Improving Automatic Language Identification Performance," Proc. Eurospeech97, Vol 1, pp 51-54, 1997.

[4] ITU-T Recommendation G.723.1, \Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3kb/s," March 1996.

[5] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech." IEEE Trans. Speech and Audio Proc., SAP-4(1), Jan. 1996, pp. 31-44

[6] Y. Yan and E. Bernard, "An approach to automatic language identification based on language-dependent phone recognition." Proc. ICASSP '95, vol. 5, May 1995, pp. 3511-3514. .

[7] Gaussian Mixture Models, Douglas Reynolds, MIT Lincoln laboratory, 244 Wood St., Lexington, MA 02140, USA

[8] D.A. Reynolds, \Comparison of Background Normalization Methods for Text-Independent Speaker Verification," Proc. Eurospeech97, Vol 1, pp 963-967, 1997.

[9] Universal Background Models, Douglas Reynolds, MIT Lincoln Laboratory, 244 Wood St., Lexington, MA 02140, USA.

[10] Alejandro Abejón González "Phonotactic speaker and language recognition"