ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 17, Issue 2, 2023

https://doi.org/10.5281/zenodo.14505515

Accurate speaker identification with Bluetooth speech transmission

Gattu Sandeep[1], Assistant Professor[1], Department of ECE, Siddhartha Institute of Technology & Sciences, Telangana, India

Vijay Koraveni[2,] Assistant Professor[2], Department of CSE, Siddhartha Institute of Technology & Sciences, Telangana, India.

Abstract

This paper studies the process of speaker identification over Bluetoothnetworks. Bluetooth channel degradations are considered prior to the speakeridentification process. The work in this paper employs Mel-frequency cepstral coefficients for feature extraction. Features are extracted from different transforms of the received speech signals such as the discrete cosine transform (DCT), signal plus DCT, discretesine transform (DST), signal plus DST, discrete wavelet transform (DWT), and signal plus DWT. A neural network classifier is used in the experiments, while the training phase uses clean speech signals and the testing phase uses degraded signals due to communication over the Bluetooth channel. A comparison is carried out between the different methods of feature extraction showing that the DCT achieves the highest recognition rates.

Introduction

Speaker identification is the process of determining the identity of a speaker automatically. This is performed in two stages. In the first stage, the features are extracted from the speech signal to represent the speaker information. In the second stage, the identification process is performed with and appropriate classifier (Chadha 2011). The speaker recognition system extracts features that can efficiently represent the speaker information from the speech signal. Depending on a chosen suitable set of features, a reference model is developed for each speaker. Finally, a matching process is performed to decide whether to accept or reject the claiming speaker (Chavan and Chougule 2012). In this paper, a study is presented for speaker identification over Bluetooth networks. This study includes the effect communication through the Bluetooth channel on the speaker recognition (El-Bendary et al. 2012). Different approaches for feature extraction are experimented; features from the signals, features from the DCT of the signals, features from the DST of the signals, features from the DWT of the signals, and hybrid features from the signals and one of their transforms (Trivedi et al. 2011).In the rest of this paper, Sect. 2 introduces the speaker identification system. Section 3 introduces the Bluetooth transmission system. Section 4 presents the experimental results followed by the discussion and conclusion in Sect. 5.

Speaker identification system

Speech signal features can be extracted with different techniques such as linear prediction coefficients (LPCs), linearpredictive cepstral coefficients (LPCCs), Perceptual Linear Predictive (PLP) analysis, and Mel-frequency cepstral coefficients (MFCCs). The MFCCs are widely used for this purpose and are adopted in this paper (Han 2006).

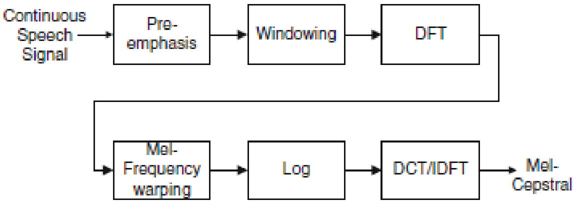


Fig. 1 MFCC extraction stages

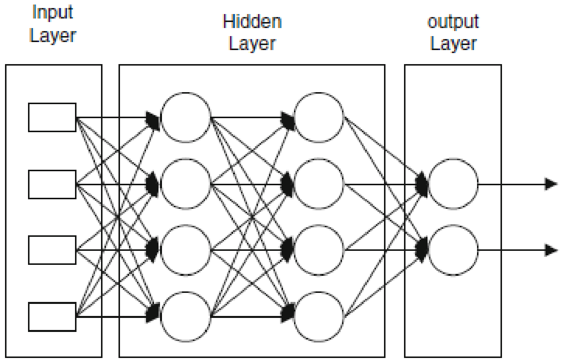


Fig. 2 The neural network structure

Mel-frequency cepstral coefficients

Mel-frequency cepstral coefficients (MFCCs) are derived through cepstral analysis,which is commonly used in speaker identification. In cepstral analysis, the excitation and vocal tract are separated to get speakerdependent information. In this analysis, the redundant pitch information is separated from the vocal tract information (Kinnunen 2003). MFCC features are also based on the human perception of frequency content, which emphasizes low frequency components more than high frequency components.

https://doi.org/10.5281/zenodo.14505515

Calculation of MFCC featuresproceeds similar to the cepstral transformation process. The input speech signal is firstly pre-emphasized with a first order Finite Impulse Response (FIR) filter, and the digitalized speech is pre-emphasized to remove glottal and lip radiation effects. The filter transfer function is given by$H(z) = 1 - az-1,\ 0.9 < a < 0.99$ (1)After that, the speech is segmented into short time frames with 50% overlap between adjacent frames. A windowingfunction is applied for each frame to increase the continuity between adjacent frames. Rectangular window and Hamming window are of the most commonly used windowing functions in speech signal processing. Hamming window with N points is given by:

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \qquad (2)$$

The FourierTransform is computed and the resulting spectrum magnitude is warped on the Mel-scale. The log of thisspectrum is then taken and a DCT is applied (Pullella and Togneri 2006). The Mel is considered as a measuring unitof perceived pitch or frequency of a tone. The Mel-scale is a mapping between the real frequency scale (Hz) and theperceived frequency scale (Mels). The Mapping is virtually linear below 1 KHz and logarithmic above according to:

$$f_{mel} = 2595 \cdot log_{10}\left(1 + \frac{f_{linear}}{700}\right) \qquad (3)$$

The MFCC extraction block diagram is shown in Fig. 1. The final stage involves performing a discrete cosine transform (DCT) on the log of the Mel spectrum. If the output of the mth Mel filter is ˜$S\ (m)$, then the MFCCs are given as:

$$c_j = \sqrt{\frac{2}{N_f}} \sum_{m=1}^{N_f} \log\left(\tilde{S}\,(m)\right) cos\left(\frac{j\pi}{N_f}(m - 0.5)\right) \qquad (4)$$

where ($j = 0, 1 \ldots J - 1$), J is the number of MFCCs, and N_f is the number of Mel filters.

Artificial neural networks

Neural Network classifiers simulate the characteristics of biological neurons within the human brain. Neural networks learn by mappings between inputs and outputs, and this is useful when the underlying statistics of the task are not well known (Galushkin 2007). Multi-layer perceptrons (MLPs) represent a popular type of neural networks, and consist of an input layer, one or more hidden layers, and one output layer as shown in Fig. 2.The inputs are fed into the input layer and get multiplied by interconnection weights as they pass from the input layer to the hidden layer. Then, they get summed and processed by a nonlinear function. Finally, the data is multiplied by an interconnection weights, and then processed for the last timewithin the output layer to produce the neural network output.Mapping is needed to train the neural network .The MLPs have been applied successfully to solve some difficult and diverse problems by training them in a supervised manner with a highly popular algorithm known as the error backpropagation algorithm (Galushkin 2007; Shuling and Wang2009) Training is performed using back-propagation, which is an iterative gradient algorithm. This algorithm is based on the error correction learning rule. Basically, error Backpropagation learning consists of two passes through the different layers of the network; forward pass and backward pass. In the forward pass, an activity pattern (input vector) is applied to the sensory nodes of the network, and its effect propagates through the network layer by layer. Finally, a set of outputs is produced as the actual response of the network. During the forward pass, the synaptic weights of the networks are all fixed. During the backward pass, on the other hand, the synaptic weights are all adjusted in accordance with an error correction rule. The actual response of the network is subtracted from a desired (target) response to produce an error signal. This error signal is then propagated backward through the network (Haykin 1999).

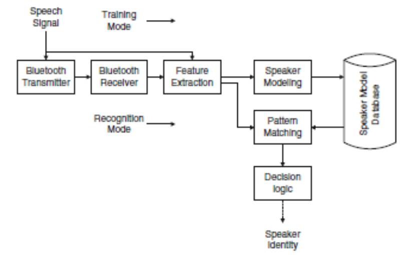


Fig. 3 Block diagram of the recognition system (tested with Bluetooth transmitted samples)

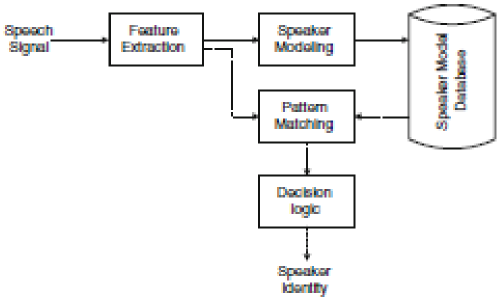


Fig. 4 Block diagram of the recognition system (tested with samples recorded directly)

Bluetooth

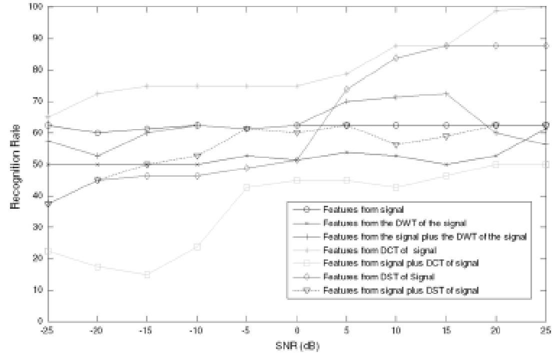
https://doi.org/10.5281/zenodo.14505515

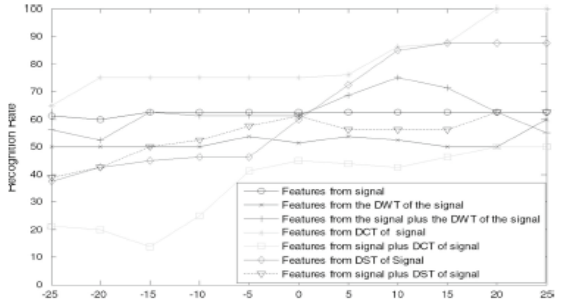
Bluetooth is a short-range wireless networking technology that allows easy interconnection ofmobile computers, mobile phones, headsets, PDAs and computer peripherals such asprinters, without the need for cables. It is designed to be of low-cost and low form-factor. So, much design work is required to optimize resource usage.

Bluetooth implementation

Bluetooth wireless technology provides solutions for interconnecting different devices like laptop computers withmobile phones. Bluetooth uses the unlicensed instrumentation, scientific, and medical (ISM) band around 2.4 GHz. It shares the channel with devices used for other applications including cordless phones, garage door openers, highway toll transponders, and outside broadcasting equipment

It is also susceptible to interference from microwave ovens, which emit radiation in this bandwidth. There are two other wireless networking standards that use the same frequency band, namely 802.11b or "Wi-Fi" and Home RF.Wi-Fi uses direct sequence spread spectrum and Home RF uses frequency hoppingDue to the limited power of the Bluetooth antenna (1mW or 0 dBm), the operating distance is limited to 0–10 m. Atypical Bluetooth connection is composed of a master an d 1 up to a maximum of 7 active slaves, which forms a Bluetooth piconet. A master initiates the connection and controls it (AlBawab et al. 2003).





Two types of connections exist, Synchronous Connection Oriented (SCO), used for 64 kbps coded speech, and Asynchronous Connectionless Link (ACL), used for other kinds of datawithamaximum rate of 723.2 kbps. The SCO link is a point-to-point link between a master and a single slave in the piconet.The master maintains theSCOlink by using reserved slots at regular intervals. The SCO link is typically used for voice connections. TheACL link is a point-to-multipoint link between the master and all the slaves participating in the piconet. In the slots not reserved for the SCO links, the master master can establish an ACL link on a per-slot basis to any slave,including the slaves already engaged in an SCO link. There is no slot reservation. For most ACL packets, packet retransmissionis applied to assure data integrity (Russo 2005).

Speech communication over Bluetooth networks

The Bluetooth standard uses frequency hopping technique to select the frequency to which members of a piconet are tuned each time a transmission takes place. It selects from 79 available channels, each of width 1 MHz, spread around 2.45 GHz, and falling in the ISM free band (2.4–2.483 GHz). Bluetooth uses Time Division Duplex (TDD), where each device is given the chance to use the channel. This prevents two or more members from transmitting at the same time and in turn prevents crosstalk from within the piconet. Interference from other piconets is possible if transmission overlaps in time and frequencyDuring a connection, radio transceivers hop from one channel to another in a pseudo-random way. Bluetooth channel is divided into 625 µs intervals called slots. Each slot uses different frequency hops. This gives a nominal hop rate of 1600 hops per second. One packet can be transmitted during the interval/slot. Subsequent slots are alternately used for transmitting and receiving, which results in a TDD scheme. Two or more Bluetooth units that share a channel form a piconet. They are synchronized to a common clock and frequency hopping pattern. One device called the master provides the synchronization. All other devices are defined as slaves (Russo 2005).

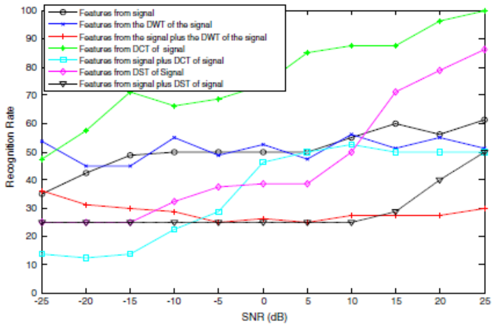


Fig. 9 Recognition rate versus SNR (dB) for noisy signals (Experiment 2)

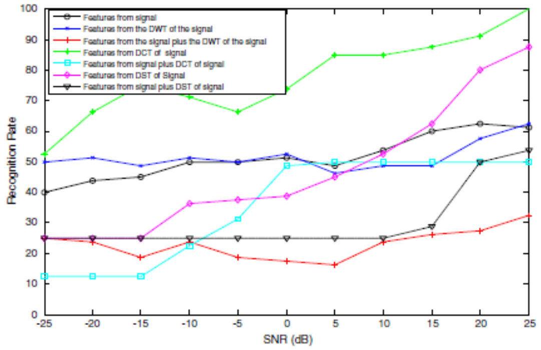


Fig. 10 Recognition rate versus SNR (dB) for wavelet denoised signals (soft thresholding) (Experiment 2)

Bluetooth deploys two techniques to correct errors, Forward Error Correction (FEC) and Automatic Repeat Request (ARQ).With a 2/3 rate FEC, a (15–10) shortened Hamming code is applied to the data payload and the resulting parity bits are transmitted with the payload. The 2/3 FEC can correct a 1-bit error and can detect a 2-bit error in every 10 bits. With ARQ, the transmitter retransmits the same packet until either a positive acknowledgement (ACK) is received or the number of retransmissions exceeds a certain threshold, which depends on the time-sensitivity of the data. Both techniques cause additional overhead in transmission, especially in error free connections (by FEC), or in bursty channels (by ARQ).

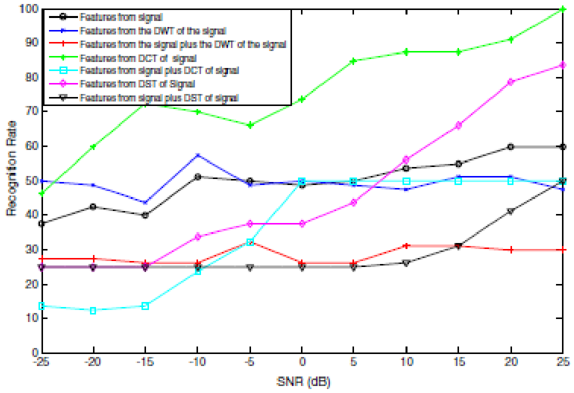


Fig. 11 Recognition rate versus SNR (dB) for wavelet denoised signals (hard thresholding) (Experiment 2)

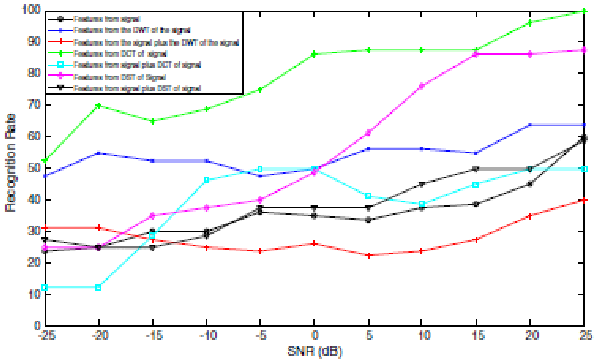


Fig. 12 Recognition rate versus SNR (dB) for low-pass filtered noisy signals (Experiment 2)

Experiments and results

A data set is generated with eight samples, four men and four women to be used in the experiments. Sound files with a “.wav” extension are sent on the transmitting end and recorded on the receiving end using Microsoft sound recorder. This sound recorder is very useful for streaming audio applicationsThe transmitting set is either sound samples recorded using wired microphone directly to the computer or samples recorded using Bluetooth handset via Bluetooth channel on the receiving end (the Bluetooth handset is located at a 1 m distance from the receiving end). The wired microphone directly recorded sounds have been used for the training phase, while testing phase has been performed twice, one with the directly recorded samples and the other with samples transmitted via Bluetooth as shown in Figs. 3 and 4.In the training phase of the experiment, the MFCCs and polynomial coefficients are estimated to form the feature vectors of the database. In the testing phase, similar features to those used in the training are extracted from the degraded or

https://doi.org/10.5281/zenodo.14505515

transmitted samples and used for matching. Seven methods for extracting features have been adopted. In the first method, the MFCCs and the polynomial coefficients are extracted from the time domain signals only. In the second method, the features are extracted from the DWT of these signals. In the third method, the features are extracted from both the original signals and the DWT of these signals and concatenated. In the fourth method, the features are extracted from the DCT of the time domain signals. In the fifth method, the features are extracted from both the original signals and the DCT of these signals and concatenated. In the sixth method, the features are extracted from the DST of the time domain signals. In the last method, the features are extracted from both the original signals and the DST of these signals and concatenated. Studying and comparing all these extraction methods for two different experiments have been considered, and the results are shown in Figs. 5, 6, 7, 8, 9, 10, 11 and 12.

Conclusions

This paper presented a robust feature extraction method from speech transmitted over Bluetooth networks. Different transforms have been investigated for robust feature extraction. Results show that the features extracted from DCT of signals recorded best recognition results in experiments. This is attributed to the energy compaction property of the DCT which allows efficient characterization of speakers with few features.

References

- *Al Bawab, Z., et al. (2003). Speech recognition over bluetooth wireless channels. EUROSPEECH 2003, Geneva.*

- *Chadha, A. (2011). Text-independent speaker recognition for low SNR environments with encryption. International Journal of Computer Applications, 31(10) (0975–8887), October 2011.*

- *Chavan, M. S., &Chougule, S. V. (2012). Speaker features and recognition techniques: A review. International Journal of Computational Engineering Research. ISSN: 2250–3005, IJCER, May–June 2012, Vol. 2, Issue No. 3, 720–728.*

- *El-Bendary, M. A. M. M. El-azm, A. E. A., El-Fishawy, N. A., Shawki, F., Abd-ElSamie, F. E., El-Tokhy, M. A. R., et al. (2012). Performance of the audio signals transmission over wireless networks with the channel interleaving considerations. EURASIP Journal on Audio, Speech, and Music Processing, 4.*

- *Galushkin, A. I. (2007). Neural network theory. Berlin, Heidelberg: Springer.*

- *Han,W. (2006). Speech recognition IC with an efficientMFCC features. The Chinese University of Hong Kong, Sept. 2006.*

- *Haykin, S. (1999). Neural networks, 2nd ed. McMaster University, Hamilton, ON, Canada.*

- *Kinnunen, T. (2003). Spectral Features for automatic text-independent speaker recognition.University of Joensuu, Department of Computer Science, Joenssuu, Finland.*

- *Pullella, D., &Togneri, R. (2006). Speaker identification using higher order spectra. University of Western Australia.*

- *Russo, M. (2005). Speech recognition over Bluetooth ACL and SCOlinks: A comparison. IEEE.*

- *Shuling, L.,&Wang, C. (2009) Nonspecific speech recognitionmethodbased on composite LVQ1 and LVQ2 network. Chinese Control andDecision Conference (CCDC), 2304–2388.*

- *Trivedi, N., Kumar, V., Singh, S., Ahuja, S., & Chadha, R. (2011).Speech recognition by wavelet analysis. International Journal ofComputer Applications, 15(8) (0975–8887) February 2011.*