Speaker Recognition

Sumanta Karmakar, Pratik Dey

Advanced Engineering

Speaker Recognition

Sumanta Karmakar and Pratik Dey

Editor: Sahadev Roy, PhD

Speaker Recognition

Sumanta Karmakar and Pratik Dey

1.1 Introduction

The objectives speaker recognition is to get hands-on experience with signal processing in both the time and the frequency domain, to become familiar with MATLAB programming, and to receive a small bit of insight into the principles of speech analysis [1]. This was accomplished by recording four speech segments from each person in the department all of them varying slightly. Comparisons and analysis were then made on each signal, dependent upon different condition [2]. These instructions involved six different procedures to apply to the recorded signals, which were speech editing, speech degradation, speech enhancement, pitch analysis, formant analysis, and waveform comparison [3]. Speech analysis was a simple cut-and-paste type procedure [4]. Speech degradation and speech enhancement were related sections, in which a signal was taken, noise was added, and then a low pass filter was used to help diminish that noise. Pitch analysis was a useful way to roughly tell if a speaker was male or female based on the average pitch derived from the pitch contour.

Progress of speaker recognition systems is based on characteristics of an individual's voice like a fingerprint reorganization system [5]. Speaker identification study continues today under the area of the field of digital signal processing (DSP). In the current design project a basic speaker identification algorithm has been written to sort through a list of files and choose the 12 most likely matches based on the average pitch of the speech utterance as well as the location of the formants in the frequency domain representation [6]. In addition, experience has been gained in basic filtering of high frequency noise signals with the use of a Butterworth filter as well as speech editing techniques.

1.2 Speech

Speech is the vocalized form of human language. It is depends upon the syntactic mixture of lexical and names that are drawn from very large [7]. Each verbal word is formed out of the phonetic permutation of a limited set of vowel and consonant units. These vocabularies, the syntax which

structure set of language sound unit differ; create the existence of many thousands of different types of mutually unintelligible human languages [8]. The vocal ability that enable human to create speech also provide humans with the ability to sing [9]. Speech is researched in terms of the speech production and speech perception of the sounds used in vocal language.

1.3 Principle of Speaker Recognition

Speaker recognition is the process identification of a individual from uniqueness of voices which is also known as voice biometrics [10]. There is a difference between speaker recognition (recognizing who is speaking) and speech recognition (recognizing what is being said) voice recognition can be used for both. Recognizing the speaker can make simpler the task of translating speech in system that have been taught on specific person's voices or it can be used to validate or verify the identity of the speaker [11]. Speaker verification has earned speaker recognition its classification as a behavioral biometric. The system that we will describe is classified as text-independent speaker identification system since its task is to identify the person who speaks regardless of what is saying.

Speaker recognition can be classified into identification and verification. Speaker identification is the processes of formative which register speaker provide a given utterance [12]. Fig 1 shows the basic structure of speaker recognition and confirmation systems.

At the highest level, all speaker recognition systems contain two main modules : feature extraction and feature matching.

Feature extraction is extracts a small amount of data from the voice signal that can soon after be used to represent each speaker. Source of conflict is the speaker himself/herself. Speech signals in training and testing sessions can be greatly different due to many facts such as people voice change with time, health.

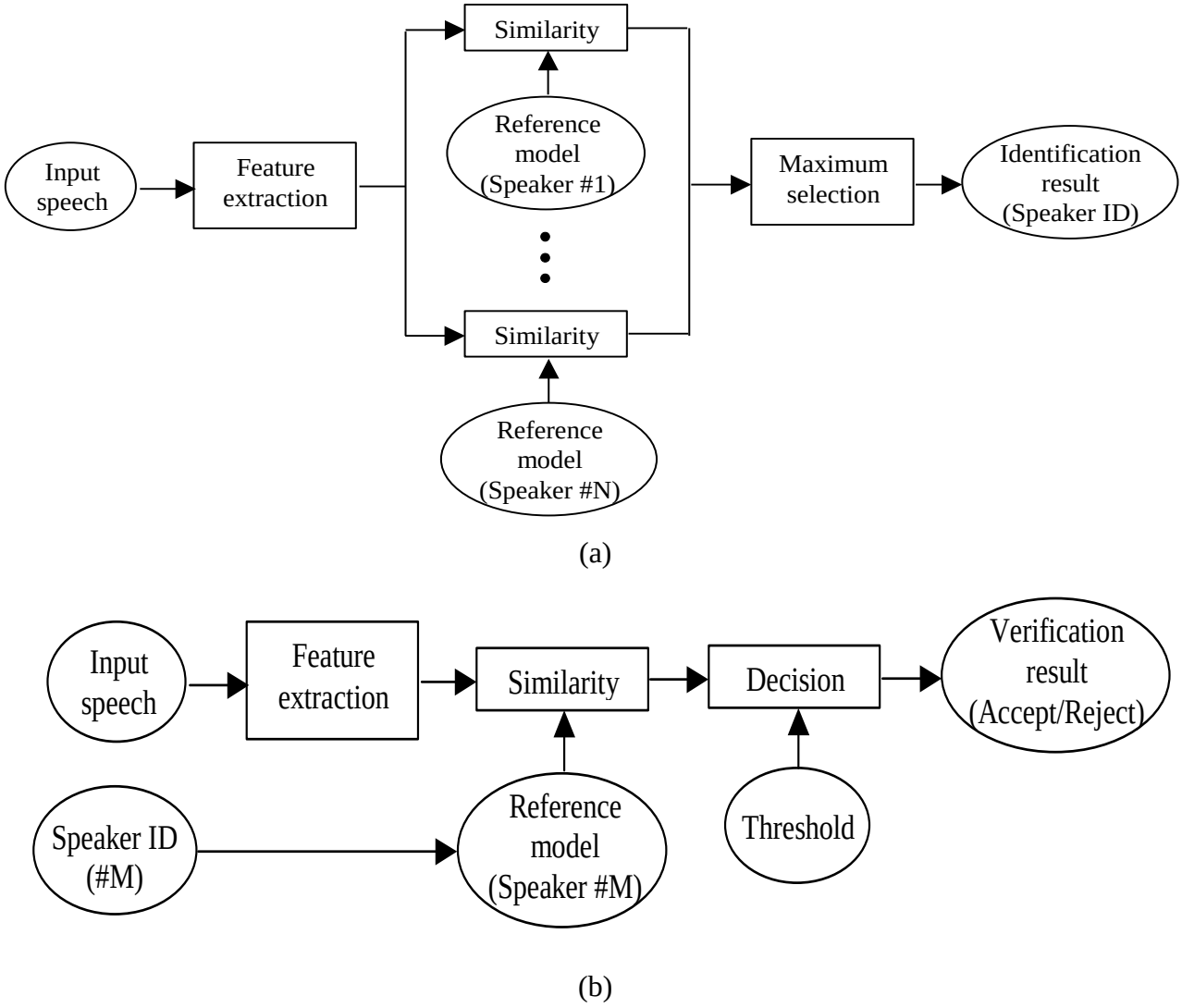

Fig 1. *Basic structures of speaker recognition systems (a) Speaker identification (b) Speaker Verification*

1.4 Analysis of Discrete-Time Speech Signals

Most speech processing applications utilize certain properties of speech signals in accomplishing their tasks. This section describes these properties or features and how to obtain them from a speech signal s(n),i.e., speech analysis. This typically requires a transformation of s(n) into a set of parameters, or more generally into a set of signals, with the purpose often being data reduction. The relevant information in speech can often be represented very efficiently [13]. Speech analysis

extracts features which are pertinent for different applications, while removing irrelevant aspects of the speech. Several important speech analysis techniques include:

•Time-frequency representation of speech signals based on a STFT analysis;

•Linear predictive speech analysis based on the characterization of all-pole digital filters;

•Cepstral analysis of speech signals based on a specific DSP technique for mixed linear and nonlinear discrete-time system analysis;

•Speech formant tracking based on analysis of digital resonant systems;

•Voicing pitch tracking based on time and frequency analysis techniques for discrete-time signals; and

•Speech analysis using auditory models based on filter-bank signal decomposition.

All these analysis techniques have significant and wide applications in speech technology and in enhancing one understand of the fundamental properties of the speech process.

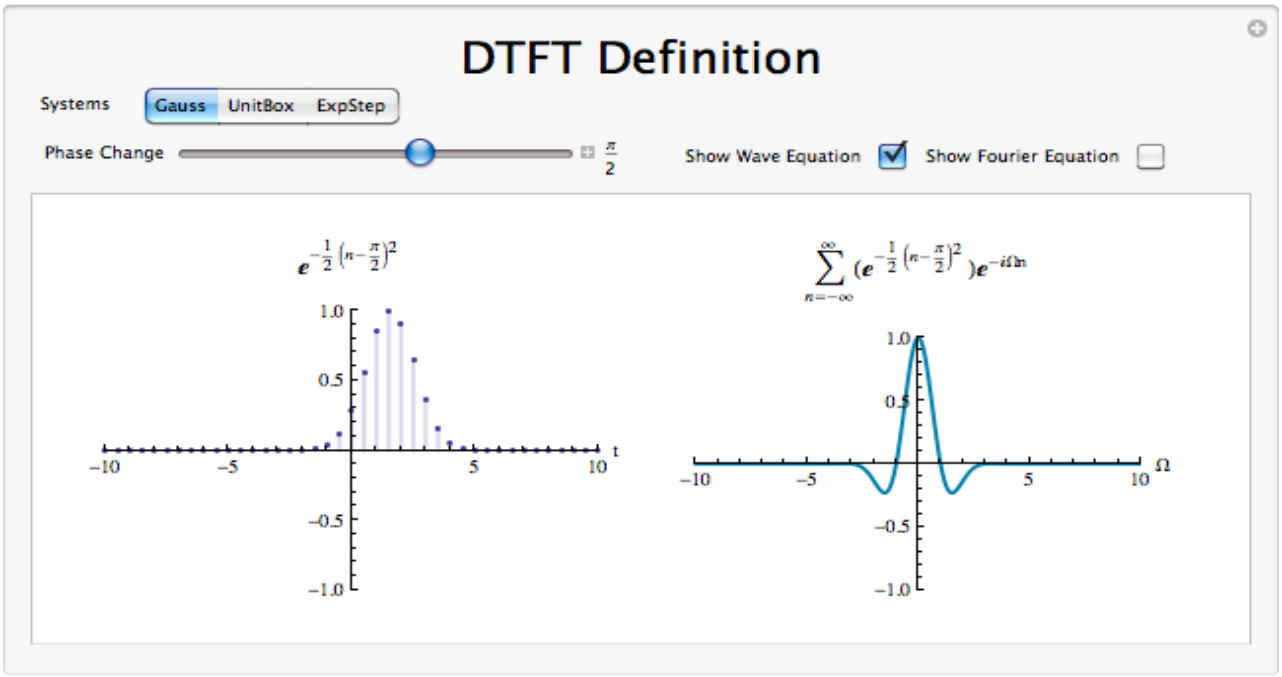

Fig 2. *DTFT waveform*

1.5 Times-Frequency Analysis of Speech

Short-time Fourier transforms (STFT), which serves as a commonly used tool for general time-frequency signal analysis. Applications of the STFT-based time-frequency analysis to speech

signals have shed crucial insight into physical properties of the speech signals. Such an analysis enables one to uncover the underlying resonance structure of speech and the vocal tract motion over time, which is responsible for the generation of the directly observable time-domain speech signal [14]. This is essentially a way of compressing a three-dimensional function into a two-dimensional display medium exploiting the visual perceptual dimension of darkness.

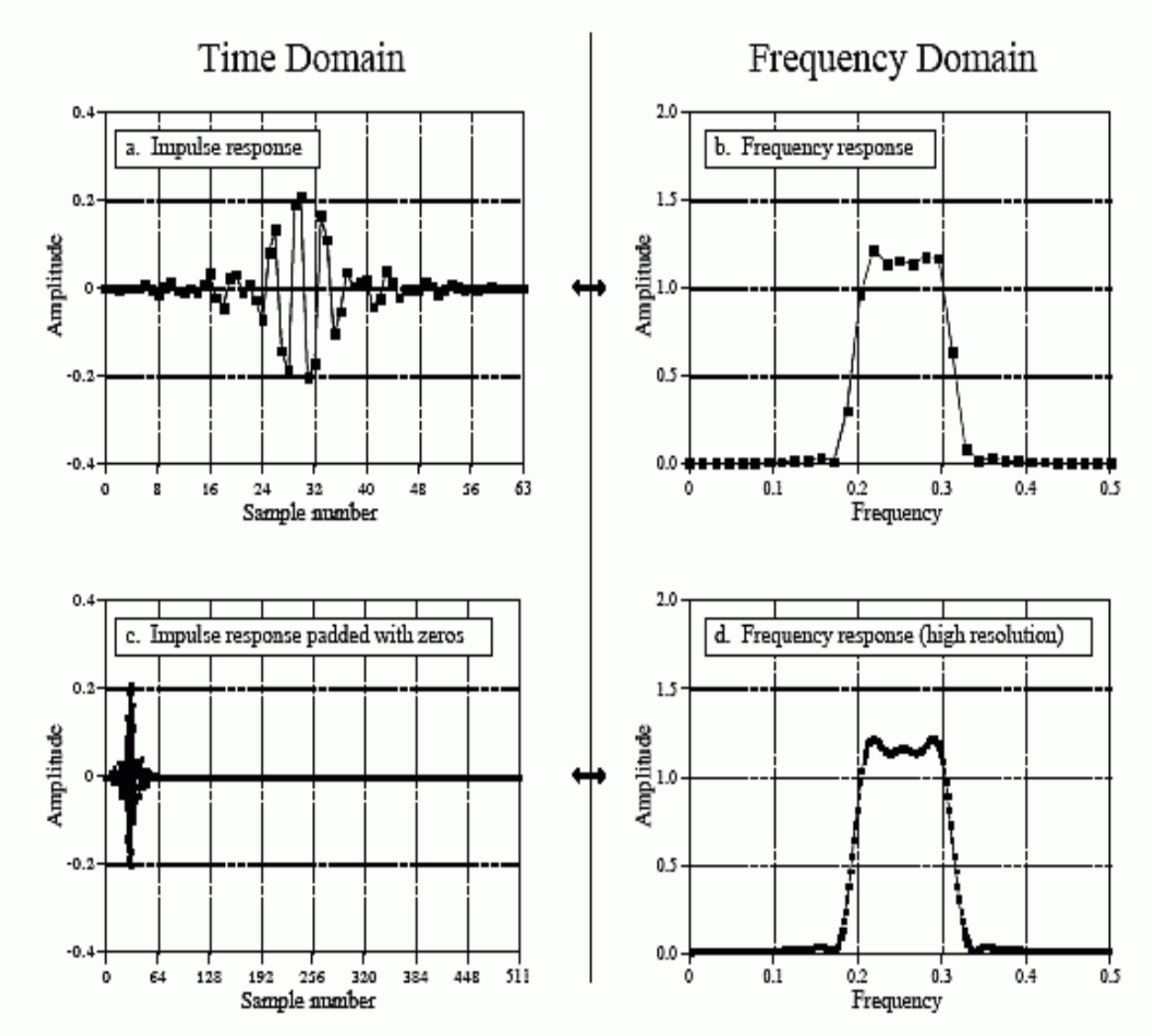

Fig 3. *Time and Frequency domain response*

1.6 Time-Domain and Frequency-Domain Properties of Speech

Analyzing speech in the time domain often requires simple calculation and interpretation in frequency domain. Most models of speech production assume a noisy or periodic waveform exciting a vocal-tract filter [15]. The excitation and filter can be described in either the time or frequency domain, but they are often more consistently and easily handled spectrally. For example, repeated utterances of the same text by a single speaker often differ significantly temporally while being very similar spectrally. Human hearing seems to pay much more attention to spectral aspects of speech (especially power distribution in frequency) than to phase or timing aspects.

1.7 Waveforms

Time-domain speech signals are also called speech waveforms. They show the acoustic signals or sounds radiated as pressure variations from the lips while articulating linguistically meaningful information. The amplitude of the speech waveform varies with time in a complicated way, including variations in the global level or intensity of the sound [16]. The probability density function of waveform amplitudes, over a long-time average, can be measured on a scale of speech level expressed as sound dB. This function has a form close to a double-sided (symmetric) exponential at high amplitudes, and is close to Gaussian at low amplitudes. The entire probability density function can be approximated by a sum of exponential and Gaussian functions. Such distributions can be exploited in speech coding and recognition.

1.8 Fundamental Frequency

Under detailed examination, a speech waveform can be typically divided into two categories:

1. a quasi-periodic part which tends to be repetitive over a brief time interval;
2. a noise-like part which is of random shape.

For the quasi-periodic portion of the speech waveform, the average period is called a fundamental period or pitch period. Its inverse is called the fundamental frequency or pitch frequency, and is abbreviated F0. (Although pitch is actually a perceptual phenomenon, and what is being measured is actually F0, we follow tradition here and consider pitch synonymous with F0.) The fundamental frequency corresponds to vocal cord vibrations for vocalic sounds of speech. F0 in a natural speech waveform usually varies slowly with time. It can be 80 Hz or lower for male adults and above 300

Hz for children and some female adults. F0 is the main acoustic cue for intonation and stress in speech, and is crucial in tone languages for phoneme identification. Many low-bit-rate voice coders require F0 estimation to reconstruct speech.

1.9 Overall Power

The overall power of the speech signal corresponds to the effective sound level of the speech waveform averaged over a long-time interval. In a quiet environment, the average power of male and female speech waveforms measured at 1 cm in front of a speaker's lips is about 58 dB. Male speech is on average about 4.5 dB louder (greater power) than female speech. Under noisy conditions, one's speech power tends to be greater than in a quiet environment (i.e., we speak more loudly in order to be heard). Further, not only the overall power and amplitude is increased, but also the details of the waveform changes in a complicated way. In noisy environments a speaker tends to exaggerate articulation in order to enhance the listener's understanding, thereby changing the spectrum and associated waveform of the speech signal.

1.10 Overall Frequency Spectrum

While the spectral contents of speech change over time, if we take the discrete-time Fourier transform (DFT) of the speech waveform over a long-time interval, we can estimate the overall frequency range that covers the principal portion of the speech power. Such information is important for the design of speech transmission systems since the bandwidth of the systems depends on the overall speech spectrum rather than on the instantaneous speech spectrum. When such an overall frequency spectrum of speech is measured in a quiet environment, it is found that the speech power is concentrated mainly at low frequencies. For example, over 80% of speech power lies below 1 kHz. Beyond 1 kHz, the overall frequency spectrum decays at a rate of about −12 dB per octave. Above 8 kHz, the speech power is negligible. In a telephone channel, due to telephone bandwidth limitations, the overall frequency spectrum becomes negligible above 3.2 kHz, losing some information, mainly for consonants. If the long-time Fourier transform analyzes only a quasi-periodic portion of the speech waveform, we will see the frequency components in harmonic relations, i.e., integer multiples of a common frequency. This common frequency, also called the lowest harmonic component, is the pitch.

1.11 Conclusion

Overall, the project went very well, and we have learned quite a lot. Speech editing is nothing more than moving about some arrays of numbers. Enhancement filters can be used to remove both natural and intentional noise, to a reasonable extent. And pitch and formant analysis can be used to give a general idea of whether two speakers are the same person or not. There are also other factors, beyond speaker variability, that present a challenge to speaker recognition technology. Examples of these are acoustical noise and variations in recording environments (e.g. speaker uses different telephone handsets).The defect, however, are obvious in the waveform comparison. While these approaches can be used to give a rough estimate or to aid in human decisions about whether two voices are the same, computer programs like these are simply not advanced enough to be completely automated and foolproof. In other words, this is not a black box where you do not have to know anything about how the program works and just expect an accurate answer based on a certain set of inputs. Other things that we would like to explore in the subject include Delta-Cepstrum coefficients and perceptual linear predictive coefficients in order to see how much they could help with or replace pitch and formant analysis. Maybe a combination of all four would give a much higher confirmation percentage.

References

[1] Quatieri, T. F. (2006). Discrete-time speech signal processing: principles and practice. Pearson Education India.

[2] Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech, and Signal Processing, 29(2), 254-272.

[3] Beek, B., Neuberg, E., & Hodge, D. (1977). An assessment of the technology of automatic speech recognition for military applications. IEEE Transactions on Acoustics, Speech, and Signal Processing, 25(4), 310-322.

[4] Chu, W. C. (2004). Speech coding algorithms: foundation and evolution of standardized coders. John Wiley & Sons.

[5] Strait, D. L., Parbery-Clark, A., Hittner, E., & Kraus, N. (2012). Musical training during early childhood enhances the neural encoding of speech in noise. Brain and language, 123(3), 191-201.

[6] Jayant, N. S. (1974). Digital coding of speech waveforms: PCM, DPCM, and DM quantizers. Proceedings of the IEEE, 62(5), 611-632.

[7] Hollien, H. (2012). About forensic phonetics. Linguistica, 52(1), 27.

[8] Harper, M. P., & Maxwell, M. (2008). Spoken language characterization. In Springer Handbook of Speech Processing (pp. 797-810). Springer Berlin Heidelberg.

[9] akobson, R., & Waugh, L. R. (2002). The sound shape of language. Walter de Gruyter.

[10] Campbell, J. P. (1997). Speaker recognition: A tutorial. Proceedings of the IEEE, 85(9), 1437-1462.

[11] Taylor, J. R., Cooren, F., Giroux, N., & Robichaud, D. (1996). The communicational basis of organization: Between the conversation and the text. Communication theory, 6(1), 1-39.

[12] Tao, T. (2008). A review of contributions by Australian research institutions into speech processing.

[13] Cotos, E. (2010). Automated writing evaluation for non-native speaker English academic writing: The case of IADE and its formative feedback. Iowa State University.

[14] Schuller, D. I. B., Ablaßmeier, D. I. M., Müller, D. I. R., & Poitschke, D. I. T. (2006). Speech communication and multimodal interfaces. In Advanced man-machine interaction (pp. 141-190). Springer Berlin Heidelberg.

[15] Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. Digital signal processing, 10(1-3), 19-41.

[16] Auckenthaler, R., Carey, M., & Lloyd-Thomas, H. (2000). Score normalization for text-independent speaker verification systems. Digital Signal Processing, 10(1), 42-54.

Authors

Sumanta Karmakar, Currently working as Assistant Professor, ECE Department at Asansol Engineering College, West Bengal, India. B.E. from Burdwan University, M.Tech from National Institute of Technical Teachers Training, Kolkata and pursuing PhD from IIT(ISM), Dhanbad. Currently involved in Rectenna related research work. He is also serve as editorial board member of International Journal of Advanced Engineering and Management (ISSN 2456-8066).
sumanta.karmakar@gmail.com

Pratik Dey, M.Tech from Asansol Engineering College, ECE Dept. (Specialization: Communication Systems). Completed B.Tech from Asansol Engineering College, ECE Dept., Asansol, West Bengal.
pratikdey87@gmail.com