Review Article Impact Factor: 4.226 ISSN: 2319-507X
Rita R. Ingle, IJPRET, 2016; Volume 4 (9): 691-700 IJPRET

INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

SPEAKER RECOGNITION IN REVERBERANT CONDITIONS: A REVIEW

RITA R. INGLE, N. P. JAWARKAR

B. N. College of Engineering, Pusad (MS), India

Accepted Date: 15/03/2016; Published Date: 01/05/2016

Abstract- The speaker recognition refers to determining the person talking from a set of known voices or speakers. There are many studies which focus on the speaker identification in clean and noisy environments. The performance of speaker identification process degrades in reverberant environments, as reverberation leads to clear physical effects on the perceived signals. Review of the research carried out in the speaker recognition under reverberant conditions is given in this paper. The paper mainly focuses on the various feature extraction methods, different speaker modelling techniques and various methods of artificial reverberation. The effect of reverberation on speech intelligence is also discussed.

Keywords: Speaker Recognition, Classification, Artificial Reverberation, MFCC

PAPER-QR CODE

Corresponding Author: MS. RITA R. INGLE

Access Online On:

www.ijpret.com

How to Cite This Article:

Rita R. Ingle, IJPRET, 2016; Volume 4 (9): 691-700

INTRODUCTION

Speech has evolved as a primary form of communication between humans. Speech carries information at several lev-els, viz. speaker specific information, the message expressed as a sequence of words or phrases, information about the acoustic environment in which the speech is re-corded and transmitted, etc. Speaker recognition is the process of extracting the features from a speech sample and identifying the person speaking. Based on the learning mode, speaker recognition can be categorized into supervised mode and un-supervised mode. Speaker identification and verification task comes under the category of supervised mode where the prior information about the speaker is known and is used in the speaker enrolment phase. Speaker diarization (also called speaker segmentation and clustering), on the other hand, comes under the category of unsupervised learning mode. The speaker identification refers to determining the person talking from a set of known voices or speakers [1]-[3]. Speaker identification can be classified into text-dependent and text-independent tasks. In the text dependent case, the utterance presented to the recognizer is known beforehand whereas, no assumptions about the text being spoken is made in text independent case. Speaker verification accepts or rejects the identify claim of a speaker.

This paper presents an overview of speaker recognition technologies in reverberant condition. The remaining of this paper is organized as follows. Section II provides a general overall overview of speaker recognition technologies. Section III presents Reverberation and IV presents the low-level and high-level features for speaker recognition. Section V presents the modelling techniques for speaker models and section VI presents the conclusions.

II. SPEAKER RECOGNITION

There are generally two phases of a speaker identification/verification system [2], first phase is called enrolment or training phase, in which a user enrols by providing voice samples to the system. The system extracts speaker-specific in-formation from the voice samples to build a voice model of the enrolled speaker. The second phase is called the classification or recognition phase, in which a test voice sample is used by the system to measure the similarity of the user's voice to the previously enrolled speaker models to make a decision. In a speaker identification task, the system measures the similar-ity of the test sample to all stored voice

Input speech

Models. In speaker verification task, the similarity is measured only to the model of the claimed identity. Fig. 1 shows a general speaker recognition system architecture.

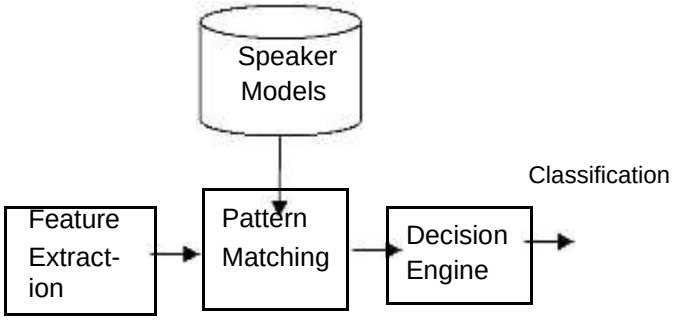

Fig. 1: A General Speaker Recognition System

III. REVERBERATION

Reverberation, in psychoacoustics and acoustics, is the persistence of sound after sound is produced [4]. A reverberation is created when a sound or signal is reflected causing a large number of reflections to build up and then decay as the sound is absorbed by the surfaces of objects in the space. Reverberation is frequency dependent: the length of the decay, or reverberation time, receives special consideration in the architectural design of spaces which need to have specific reverberation times to achieve optimum performance for their intended activity. In comparison to a distinct echo that is a minimum of 50 to 100 ms after the initial sound, reverberation is the occurrence of reflections that arrive in less than approximately 50 ms. As time passes, the amplitude of the re-flections is reduced until it is reduced to zero. The first artificial reverberation algorithms were proposed in the early 1960s, subsequently new, improved algorithms were published.

There are many applications of digital reverberation technology [4] such as room acoustic enhancement, to externalize the sound image in headphone audio, up mixing process to play stereo audio signals over multiple loudspeakers, speech processing research to evaluate its effect on speech intelligence, simulation of musical instrument virtual reality, gaming, and computer aided design of concert halls, etc.

The artificial reverberation methods can be categorized into Analog- and Digital-methods [4]. Some of the analog methods are: Using echo chambers [5], spring reverberator [6], Plate reverberator [7], Tap based techniques [8], Bucket bridge device [9].

Reverberation algorithms generally fall into one of the three categories [4]:

Delay networks - The input signal is delayed, filtered and fed back along a number of paths according to parameter-ized reverberation characteristics;

Convolution - Input signal is sim-ply convolved with estimated impulse re-sponse of an acoustic space;

Computational acoustic - Input signal drives a simulation of acoustic en-ergy propagation in the modelled geome-try.

Some of the digital artificial reverberation methods are Comb filter [10], All pass filter [10], Digital Wave Guide Net-work [11], Feedback delay network [12], Time varying reverb algorithm [13],

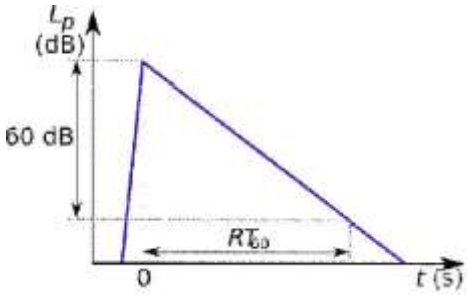

Fig.2. Sound level in a reverberant cavity ex-cited by a pulse, as a function of time.

Pseudo-Random Late Reverb Algorithm [14], Different room acoustic modelling techniques such as Finite Difference time methods [15] and Adaptive Rectangular Decomposition [16] and Convolution Techniques (equivalent to FIR filtering) [17].

Reverberation time: The time it takes for a signal to drop by 60dB is the reverberation time. RT60 is the time required for reflections of a direct sound to decay 60 dB. Reverberation time is frequently stated as a single value, if measured as a wide band signal (20 Hz to 20 kHz).

The performance of speaker recognition can be affected by noise and reverberation and hence degraded. Reverberation causes coloration of the speech signals and tem-poral spreading, which severely degrades the performance of most automated speaker recognition [18]. Some useful pa-pers which describe the effect of rever-beration on speech recognition are [19]-[20].

IV. FEATURE EXTRACTION

Speech signal includes many features that are useful for speaker discrimination. Humans rely on such different types or levels of information in the speech signal to recognize others. We can roughly cate-gorize these features into a hierarchy run-ning from low-level features to high-level features. The low level features mainly include MFCC [21], Linear

Predictive Co-efficients (LPC) [22], Modified Group De-lay features [23], Gamatone Frequency Cepstral Coefficients (GFCC) [24], Per-ceptual Linear Prediction (PLP) [25] , Wavelet based features [26], TESBCC [27].

High-level features are generally re-lated to a speaker's learned habits and style, such as particular word usage or idiolect. The low-level features are still the most popular speaker features in current state-of-the-art speaker recognition sys-tems. In this section, overview of some selected low-level features is given.

A. Mel Frequency Cepstral Coefficients (MFCC):

Mel Frequency Cepstral Coefficients (MFCCs) have been the most popular low-level features for speaker recognition and speech recognition systems. The Mel-Frequency Cepstrum (MFC) is a representation of the short-term power spectrum of a signal, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. The difference be-tween the normal cepstrum and the mel-frequency cepstrum is that in the MFCC, the frequency bands are equally spaced on the mel scale, which approximates the hu-man auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. Ac-cording to psychophysical studies, human perception of the frequency content of sound follows a subjectively defined nonlinear scale called Mel scale. Mel frequency is defined as,

$$= 2595 \qquad 1 + \frac{\qquad}{700} \qquad (1)$$

Where fmel is Mel frequency and f is the frequency in Hz. This led to the definition of MFCC. The relationship between the normal frequency and Mel-frequency is shown

Fig. 2 shows the computation flow chart of MFCCs. Take the Fourier transform of a signal in a window.

Map the power/magnitude spectrum obtained above onto the Mel scale, using triangular overlapping windows.

Take the logs of the energy at each of the mel frequencies.

Take the discrete cosine transform of log energies.

Temporal energy sub-band cepstral coefficients (TESBCC).

Sen and Basu [27] have proposed a set of parallel linearly spaced Nyquist filters for extracting TESBCC feature. The Fou-rier transform of the proposed Nyquist window function is:

$$\cos(\ ),-\ 2\quad \leq \leq\ 2 \qquad (2)$$

$$=\quad 0,\quad >\ 2$$

Speaker Recognition in Reverberant Conditions: A Review Page 3

Steps involved in computation of TESBCC are as under.

The speech signal is pre-emphasized with pre-emphasis factor and then passed through a bank of parallel fil-ters described above.

Log energy of the sub-band signal of each frame is computed.

Discrete cosine transform of log-energies in each frame is finally obtained.

V. SPEAKER MODELLING

TECHNIQUES

Using feature extraction, a spoken utterance can be represented with feature vectors. The speech signal of a person will have similar still differently arranged feature vectors. To recognize these feature vectors voice modelling is done using classifier algorithms by which a template of features is generated for a particular registered user and that is used as a reference in recognition process. It means every registered user will have a reference model in database and if a new user comes then it will be declared as unregistered one.

The classifiers used for the speaker recognition task mainly include Vector Quantization (VQ) [28], Gaussian mixture model (GMM) [29], Dynamic Time Wrapping [30], Hidden Markov Model (HMM) [31], Probabilistic Neural Network [32], Polynomial Classifier [33], VQ based PNN (34) and Support Vector Machine [35]. The generative models such as GMM and VQ estimate the feature distribution within each speaker independently. The GMM and the HMM are the most popular stochastic models for text-independent and text-dependent speaker recognition, respectively. In open-set applications (speaker Speech Pre-emphasis Framing Windowing Verification and open-set speaker identification), the estimated features can also be compared to a model that represents the unknown speakers. In a verification task, the pattern matching module outputs a similarity score between the test sample and the claimed identity. In an identification task, it outputs similarity scores for all stored speaker models.

The vector quantization is a classical quantization technique that allows the modelling of probability density functions by the distribution of prototype vectors. It was originally used for data compression. It works by dividing a large set of points (vectors) into groups having approximately the same number of points closest to them. The GMM is a density estimator. The distribution of the feature vector x is modelled clearly using a mixture of M Gaussians. They model the probability density function of observed variables using a multivariate Gaussian mixture density. Expectation maximization algorithm is used to estimate mean, covariance parameters. During recognition, a sequence of features is extracted from the input signal. Then the distance of the given sequence from the model is obtained by computing the log likelihood of given sequence. The model that provides the highest likelihood score is verified as the identity of the speaker. Table-1 shows the in-formation about the features, modelling techniques, and the reverberation methods used for speaker recognition under reverberation conditions by some of the re-searchers.

VI. **CONCLUSION**

An overview of speaker recognition in reverberant conditions are given this pa-per. This paper mainly describes methods for artificial reverberation, various features that can be used for speaker recognition and speaker modelling techniques. Reverberation causes coloration of the speech signals and temporal spreading, which de-grades the performance of automatic speaker recognition system. It is proposed to study the performance of the system in reverberant condition by using different statistical modelling techniques and features as discussed above.

REFERENCES

1. S. Furui, “Recent Advances in Speaker Recognition,” Pattern Recognition Letters, 18, (1997), 859-872.
2. J. P. Campbell, “Speaker Recognition: A Tuto-rial,” Proceedings of the IEEE, 85, 9, (1997), 1437-1462.
3. S. Tranter, and D.A. Reynolds, “An overview of automatic speaker diarization systems,” IEEE Trans. Audio, Speech and Language Processing 14, 5 (September 2006), 1557– 1565.
4. V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith and J. S. Abel, “Fifty Years of Artificial Reverberation”, Audio, Speech, and Language Processing, IEEE., vol. 20(5), pp.1421-1448 (2012).
5. B. Swedien and Q. Jones, Make Mine Music.. Winona, MN: Hal Leonard Corp., 2009.
6. L. Hammond, “Electrical musical instrument,” U.S. Patent 2,230,836, Feb. 1941.
7. B. Blesser and L. Salter, Spaces Speak, Are You Listening? Cambridge, MA: MIT Press, 2006.

8. S. Arnardottir, J. Abel, and J. Smith, “A digital model of the Echoplex tape delay,” in Proc. 125th Audio Eng. Soc. Conv., San Francisco, CA,May 2008, paper no. 7649.
9. S. Arnardottir, J. Abel, and J. Smith, “A digi-tal model of the Echoplex M. R. Schroeder and B. F. Logan, “Color-less artificial reverberation,” J. Audio Eng. Soc., vol. 9, no. 3, pp. 192–197, Jul. 1961.a
10. J. O. Smith, “A new approach to digital rever-beration using closed waveguide networks,” in Proc. Int. Comput. Music Conf., Vancouver, BC, Canada, Aug. 1985, pp. 47–53.
11. J. O. Smith, Physical Audio Signal Processing, Dec. 2010 [Online]. Available: https://ccrma.stanford.edu/~jos/pasp/, online book
12. D. Griesinger, “Practical processors and pro-grams for digital reverberation,” in Proc. AES 7th Int. Conf., Toronto, ON, Canada, May 1989, pp. 187–195.
13. P. Rubak and L. G. Johansen, “Artificial re-verberation based on a pseudo-random impulse response, part I,” in Proc. 104th Audio Eng. Soc. Conv., Amsterdam, The Netherlands, May 1998, paper no. 4725.
14. L. Savioja, T. Rinne, and T. Takala, “Simula-tion of room acoustics with a 3-D finite differ-ence mesh,” in Proc. Int. Comput. Music Conf., Aarhus, Denmark, Sep. 1994, pp. 463–466.
15. N. Raghuvanshi, C. Lauterbach, A. Chandak, D. Manocha, and M. Lin, “Real-time sound synthesis and propagation for games,” Com-mun. ACM, vol. 50, no. 7, pp. 66–73, Jul. 2007.
16. D. Griesinger, “Beyond MLS—Occupied hall measurement with FF Ttechniques,” in Proc. 101st Conv. Audio Eng. Soc., Los Angeles, CA, Nov. 1996, preprint no. 4403.
17. S. Van Vuuren, “Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch,” in Speaker Recognition in Reverberant Conditions: A Review Page 5
18. Proc. Of Fourth International Conference on Spoken Language, 1996, pp.1788-1791.
19. O. Hazrati& P. C. Loizou, “The combined effects of reverberation and noise on speech in-telligibility by Cochlear Implant listeners,”
20. Department of Electrical Engineering, The University of Texas at Dallas, Richardson, Texas, USA, vol. 51(6), pp. 437-443, (2012).
21. A. Sehr and W. Kellermann, “Strategies for modeling reverberant speech in the feature domain”, Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg Cauerstr, Germany, ICASSP (2009). IEEE International Conference on, pp. 3725-3728. (2009).
22. S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sen-tences,” IEEE Trans Acoustics, Speech, Signal Process, vol. 28, no. 4, pp. 357–366, Aug. 1980.

23. B. S. Atal, “Effectiveness of linear prediction of the speech wave for automatic speaker iden-tification and verification,” J. Acoustical Soci-ety of America, vol. 55, no. 6, pp. 1304–1312, June 1974.
24. H. Hermansky, “Perceptual linear prediction (PLP) analysis for speech,” J. Acoust. Soc. America, vol. 82, pp. 1738–1752, Apr. 1990.
25. R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, “Application of the modified group de-lay function to speaker identification and dis-crimination,” in Proc. IEEE Int. Conf. Acous-tic, Speech, Signal Process., Montreal, QC, Canada, May 2004, vol. 1, pp. 517–520.
26. Xiaojia Zhao, Yang Shao, and DeLiang Wang, “CASA-based robust speaker identification,” IEEE Trans. Audio, Speech, Lang. Process., vol.20, no. 5, pp.1608–1616, July. 2012.
27. C. T. Hsieh, E. Lai and Y. C. Wang, “Robust speech features based on wavelet transform with application to speaker identification”, IEE 2002.
28. N. Sen and T. K. Basu, “Temporal Energy and Correlation Features from Nyquist Filter Bank for Text-Independent Speaker Identification,” in Proc. of IEEE Students' Technology Sympo-sium, IIT Kharagpur, India, 2011, pp.166–170
29. Y. Linde, A. Buzo and M. Gray, “An Algo-rithm for vector Quantization,” IEEE Trans. on Communication, vol. COM-28, no. 1, pp. 84– 95, Jan.1980.
30. S. Furui, “Cepstral analysis technique for automatic speaker verification,” IEEE Trans. Acoustic, Speech Signal Process., vol. 29, no. 4, pp. 254–272, 1981.
31. D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture models,” IEEE Trans. on Speech & Audio Processing, vol. 3, no. 1, pp. 72–83, Jan. 1995.
32. L. R. Rabiner, “A tutorial on Hidden Markov models and selected applications in speech recognition,” in Proc. IEEE, 1989, vol. 77, no. 2, pp. 257–286.
33. T. D. Ganchera, D. K. Tasoulisb, M. N. Vra-hatisb, and N. D. Fakotakisa, “Generalised lo-cally recurrent probabilistic neural networks with application to text-independent speaker verification,” Neuro Computing, vol. 70, no.7-9, pp. 1424–1438, Mar. 2007.
34. W. M. Campbell, K.T. Assaleh, and C. C. Broun, “Speaker recognition with polynomial classifiers,” IEEE Trans. on Speech & Audio Processing, vol. 10, no. 4, pp. 205–212, May 2002.
35. W. Campbell, J. Campbell, D. Reynolds, and E. Singer, “Support vector machines for speaker and language recognition,” Computer Speech Lang., vol. 20 (2-3), pp. 210–229, 2006.
36. N. P. Jawarkar, R. S. Holambe and T. K. Basu “On the use of classifiers for speaker identification,“ IEEE International Conference on Automation, Control, Energy and Systems-2014, pp. 1-6, Feb. 2014.
37. A. R. Abu-El-Quran," Talker Identification using reverberation sensing system,", IEEE Sensor Conference-2007, pp. 970-973, 2007.

38. A. Akula, V. R. Apsingekar and P. L. D. Leon, “Speaker Identification in Room Reverberation using GMM-UBM.” Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, 2009.

39. Garcia-Romero, D. Zhou, & Espy-Wilson, C. Y., “Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition.” In Acoustics, Speech and Signal Processing (ICASSP), (2012) IEEE International Conference.

40. X.Zhao, Y.Wang and D.Wang, “Robust Speaker Identification In Noisy And Reverber-ant Conditions” IEEE ICASSP, 4-9 May 2014.

41. Khamis A. Al-Karawi, Ahmed H. “Automatic Speaker Recognition system In Adverse Con-ditions – Implication of Noise and Reverbera-tion on System Performance.” International Journal of Information and Electronics Engg, no. 6, pp. 423-427, 2015.