Principle and Applications of Speaker Recognition Security System

Conference Paper · June 2018

CITATIONS
0

READS
875

3 authors:

Nilu Singh
Babu Banarasi Das University
67 PUBLICATIONS **124** CITATIONS
SEE PROFILE

Alka Agrawal
Babasaheb Bhimrao Ambedkar University
86 PUBLICATIONS **216** CITATIONS
SEE PROFILE

Prof. Raees Ahmad Khan
Babasaheb Bhimrao Ambedkar University
168 PUBLICATIONS **684** CITATIONS
SEE PROFILE

Principle and Applications of Speaker Recognition Security System

Nilu Singh[1], Alka Agrawal[2] and R. A. Khan[3]
Chennupati Jagadish[1], Chris Bailey[2], and Parviz Famouri[3]
[1,2,3]SIST-DIT, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow, UP, India

Email: nilu.chouhan@hotmail.com

***Abstract* — This paper overviews the principle and applications of speaker recognition. Speech is a natural way to convey information by humans. Speech signal is enriched with information of the individual. To recognize a person by his/her voice is known as speaker recognition (SR). Since human voice has some measurable characteristics hence it falls in the category of biometric. The term biometric is used to measure human's body related characteristics. Biometric is also known as realistic authentication. Voice biometric or speaker recognition is used to recognize an individual through their voice's individual characteristic.**

Index Terms – Speaker Recognition, Speaker Verification, Speaker Identification Open-set & Closed-set, Applications of Speaker Recognition.

I. INTRODUCTION

Human speech is a medium for expressing their thoughts during communication. Voice is a medium for human to express their thoughts and information. A speech signal is a complex signal which is packed with several knowledge resources such as acoustic, articulatory, semantics, linguistic and many more [1-2]. During communication, human easily understand information such as emotion, language, and mental status etc. This ability of human to decode information motivated researchers to cognize speech signal related information. And this idea helps to emerging a system which able to procure and process the assembled information of a speech signal. A person's voice is different from another due to the acoustic properties of speech signal. Speaker's voice is unique to an individual due to differences which occur as structure of glottal and dissimilarities in the vocal tract and the cultured speaking behaviors of individuals [1-3].

In this digital era speaker recognition is the most useful biometric recognition technique [3]. Now days many organizations like bank, industries, access control systems etc. are using this technology for providing greater security to their vast databases [3-4]. Speaker recognition is mainly divided into speaker identification (1: N matching) and speaker verification (1:1 matching). Identification is considered as more difficult than verification [5]. This is intuitive that performance of speaker identification system affected by the number of enrolled speakers (the probability of incorrect decision increases). While the performance of speaker verification system is not affected by increase in voice database size since only two speakers are compared.

In last few years, requirement for authentication has been increased with the increasing digital world of information. It has already been proved that a biometrics authentication technique increases security levels. Speaker identification is a procedure of recognizing an utterance from the enrolled speakers while speaker verification is a binary task that is either speaker accepted or rejected. Speaker verification systems are the real example of biometric authentication systems. Further, it can be categorized as text-dependent and text-independent [6]. The text-dependent systems are based on same utterance spoken by speaker in both cases i.e. training and testing while in text-independent systems it is not required to utter the same sentence/words during training and testing [5]. It is accepted that text-dependent systems provide more precise results as both the content and voice can be matched that is speaker utters exact the sentence which he/she uttered during training. While text-independent recognition systems, either use the same voice sample or different for verification/identification [7-8].

Gaussian mixture model (GMM) is one of the most common method used to speaker modeling for speaker identification. GMM is used as two distinct ways for identification system; firstly, when training database principle is the maximum likelihood (ML) and parameter estimation is performed by using expectation maximization (EM) algorithm; and secondly when the training database principle is maximum a posteriori (MAP)[9-11].

Speaker verification is used for those applications where speech is used as main component to authorize the speaker. The task of speaker identification is to decide that a given utterance comes from a certain registered speaker. Speaker verification usability is more than speaker identification. With the increase in voice database, difficulty of speaker identification increases. Speaker verification is independent of voice database population since it works only on binary decision that is acceptance or rejection. Speaker recognition system performance (recognition accuracy) is most affected by intersession variability (variability over time) and spectra of a speakers speech signal [7] [12].

II. SPEAKER RECOGNITION

It is well accepted that in this electronic era people interact using voice with the help of electronic devices. Human voice is a signal which contains several information related to human characteristics, such as emotion, words, language, speaker identity etc.. To identify a human being by their voice, required speech features are selected from speech signal with available feature extraction techniques [13]. It is the process of recognizing a person on the basis of his/her voice. Automatic speaker recognition system is categorized as speaker identification and speaker verification. Speaker verification system decision is binary i.e. 0 or 1 (accept or reject) as this justifies an identity claimed by the speaker. Speaker identification decision requires N matching and then the decision is made about acceptance or rejection [14]. Further it is distinguished as text-dependent and text-independent speaker recognition. In the text-dependent system, the recognition of speaker's identity is depends on the specific words or sentences. In the case of text-independent recognition, speakers have no restriction to speak sentence or phrases [3] [15-19]. Fig. 1, presents an overview of speaker recognition system and Fig. 2, shows the basic concept of speaker identification and speaker verification methodology.

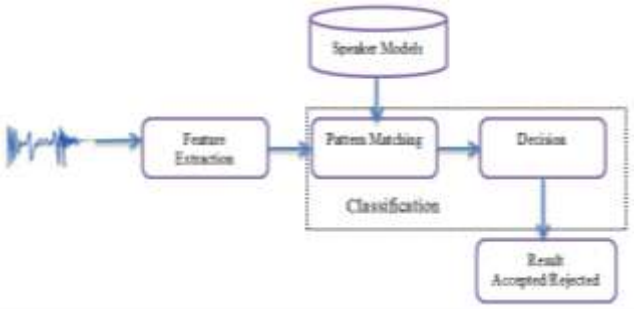


Fig. 1 Basic speaker recognition system

The popularity of speaker recognition system is due to its low cost of implementation. This is because of the easily availability of microphones and the universal telephone network. As in this digital era it is very easy to capture someone's voice and authenticate it by using speaker recognition system. The only cost is due to the software which is used for speaker recognition system. The study of speech signal is about its characteristics which distinguish one speech signal with another [3] [15] [17].

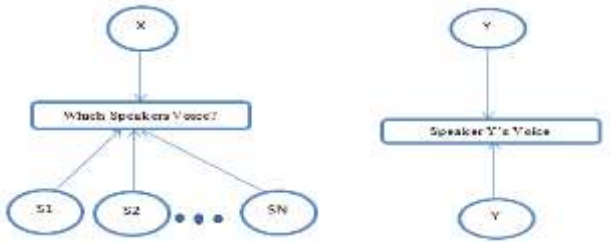


Fig. 2 Speaker identification and speaker verification

A. Speaker Identification

Speaker identification system is 1: n matching system. In this, user needs not to provide his/her identity to the system. During identification user has to input his/her speech to the ASR system and the system now decides the identity of user on the basis of the match score. In this case system has to perform N comparison (N is the number of stored speaker/user model of voice database). During identification, comparison with each registered model will produce a likelihood score, on the basis of this score higher likely model is identified for the speaker [7] [20].

From the study it is clear that speaker identification is complex than speaker verification. Hence in case of speaker identification, system performance degrades as compared to speaker verification [21]. Fig. 3, shows the process of speaker identification system.

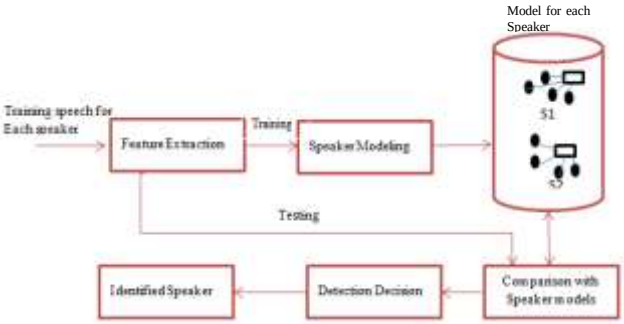


Fig. 3 Process of speaker identification system

B. Speaker Verification

Speaker verification system is 1:1 that is the system either accepts or rejects. In verification process, firstly user need to provide his/her identity and then it is checked by the system and decisions are made accordingly that whether the claimed identity is true (accept) or false (reject) [22]. Speaker verification can be explained easily with the help of an example of Automated Teller Machine (ATM). Before any transaction, users are first needed to insert the ATM card in the machine. This credit/debit card contains the information about user such as name, signature etc. Now, if the ATM is working on ASR technology then it will check that the card is used by its genuine holder by asking to produce his/her voice. Since the user has already provided his/her identity to the system, only 'yes or no' decision has to be made by the ATM machine i.e. either the card holder is accepted or rejected. This decision is made on the basis of comparing the voice input to the previous voice input provided by the user [2] [5]. Fig. 4 shows the process of speaker verification system.

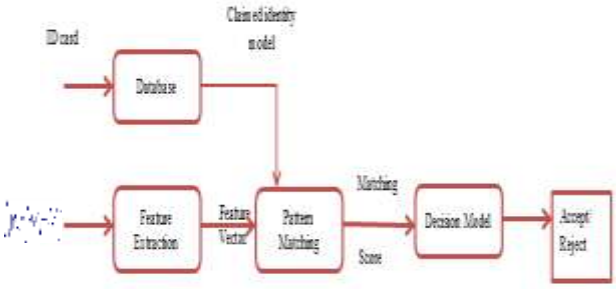
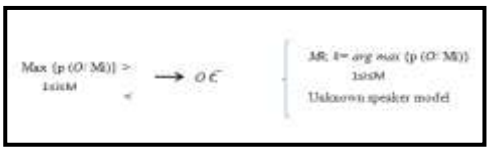


Fig. 4 Process of speaker verification system

C. Open-Set vs. Closed-Set

Speaker identification is categorized as closed-set and open-set. Open-set speaker identification for a test utterance of enrolled speakers is a twofold problem. Primarily, it is necessary to find speaker model which have best matches in the given set; Furthermore, it must be determined that the best match test utterance is actually formed by the best matched model speaker or some unknown speaker [2] [21] [23-24]. Fig. 5 shows the classification of speaker recognition system.

The possible error and problems in open set speaker identification can be examined as follows. Suppose that N speakers are enrolled in the system for voice database and M1, M2……Mi, are their statistical models [25]. If O represents the feature vector sequence extracted from the given utterance, then the open-set identification is given as follows:



Where θ is a pre-defined threshold and O is assigned to speaker model. If the maximum likelihood score is greater than the threshold θ, it means the voice is originated from a known speaker. For a given O following errors are possible [2] [23][26].

- False Acceptance (FA): The recognition system accepts a pretender as one of the registered speaker [2].
- False Rejection (FR): The recognition system discards a registered speaker [2].
- Speaker Confusion (SC): The system properly accepts a registered speaker but demented with another registered speaker [2].

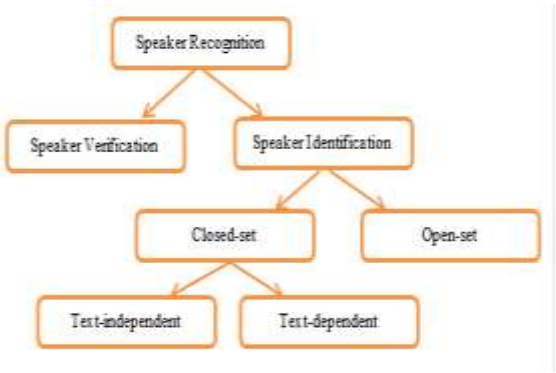


Fig. 5 Classification of speaker recognition

III. EXTRACTION OF SPEECH FEATURES

A speech signal has several features such as phonetic, prosodic and acoustic etc. Selection of the required speech features among several features is the important task. By selecting more informative speech features will help to improve system performance. Selection of less useful or not useful features for a particular task is unfavorable to the system performance. Hence, examining the new speech features for a specific task has been always an important and difficult task [27]. Speech signals do not only carry language information but it also includes key paralinguistic components, prosody. Speech features stated as prosodic features define prosody of a speech while the features which are not used to describe prosody are called acoustic features of speech. Fundamental frequency is the main component of a speech signal which is further explained as:

Fundamental Frequency features: Fundamental frequency (F0) features are useful for tonal languages. Tones are related with dynamics of the fundamental frequency. To extract F0 there are many methods used. One common method is through autocorrelation i.e. autocorrelation of the signal within a frame. In this computation, second highest peak of autocorrelation is represented as fundamental frequency of speech signal. For better accuracy or to make system more robust against noise another technique is required. It is based on observation such as tracking the peak of the autocorrelation across frames or normalizing the autocorrelation according to the analysis window [15] [28-29] [30].

Measurement unit of F0 is Hertz (Hz) i.e. number of periods within one second. Fundamental frequency period can be further divided as jitter and shimmer.

Jitter: It is frequency stability in terms of equality of period's duration. It is computed as average absolute difference between consecutive periods (divided by the average period) [31-32].

Shimmer: It is the measurement for amplitude stability of F0 periods. It is computed as Average absolute difference between the amplitudes of sequential periods (divided by the average amplitude) [31-32].

Both jitter and shimmer have their observation based thresholds and basically used in speech pathology research. From the fundamental frequency and glottal cycle point of view, there are many phonation types voice such as the following [15]:

Normal Voice: Vocal cords are in their natural mode [15] [16].

Creaky Voice, Vocal Fry: Creaky phonation is characteristically related with aperiodic glottal pulses. In such type of voice, degree of aperiodicity in the glottal source is quantified by measurement of the jitter. During creaky phonation, jitter values are higher than other phonation types [14-16].

Falsetto voice: in such type of voice production vocal folds are stretched longitudinally (thin). Therefore vibrating mass is smaller hence tone is higher [15].

Breathy voice: It is noticeable as compound phonation type. Such type of voice production has vocal fold vibration which is inefficient due to incomplete closure of the glottis [15].

F0 rises when the vowel follows an unvoiced volatile and decrease when it follows a voiced explosive.

IV. APPLICATION OF SPEAKER RECOGNITION

In the last few years use of biometric system has become a reality. There are lots of commercial as well as personal applications where biometric is used for security purpose. Speaker verification has gained a huge acceptance in both government and financial sectors for secure authentication [5] [33-34]. Australian Government organization Centrelink, use speaker verification for authentication of recipients using telephone transactions [35]. Possible applications of speaker recognition are forensic investigation, telephone banking, access control, user authentication etc. [36].

Speaker recognition have more potential than other biometrics such as face recognition, finger prints, and retina scans. The main advantage of speaker recognition over other biometric is low cost, high acceptance and non-invasive character of speech acquisition. To develop a speaker recognition system, expensive equipment as well as direct participation of speakers is not required. Speaker recognition have potential to eliminate the need of carrying debit card, credit card, remembering password for bank account or any other security locks and many other online services [6] [37-38]. With the continuous improvement in reliability of speaker recognition technology, its usability has increased. Now days, use of speaker recognition has become a commercial reality and part of consumer's everyday life [34] [39].

The performance of speaker recognition system is vulnerable to change in speaker characteristics such as age, health problems, speaking environment etc. Another disadvantage is that it is possible to play a recorded voice instead of the actual voice of a speaker [2] [34-35].

- ***Access Control:*** Controlling access to computer networks
- ***Transaction Verification:*** For telephone banking and account access control
- ***Law Enforcement:*** Used in home parole monitoring and residential call observing
- ***Speech Data Organization:*** Used for voice mail. E.g. speech skimming or audio mining applications.
- ***Personalization:*** Device customization, store and fetch personal information based on user verification for multi user device [33].

All the above mentioned applications require robust speaker recognition techniques. The requirement of robustness in case of speaker recognition system can be explained with the help of an example. In telephone based services a user speaks in many circumstances (in noisy environment or street), use different communication medium (telephone or mobile), differs the distance of microphone etc. Therefore robustness is the key factor for deciding the success of speaker recognition system. In the area, in 1983, the first international patent was filed by Michele Cavazza, Alberto Ciaramella. This invention relates to analysis of speech characteristics of speakers, in particular, to a device for a speaker's verification [40].

Barclays Wealth and Investment Management was the first organization in the world to deploy passive voice security services. The basic aim behind using voice based security was transforming the customer service experience. By using this technique customers are automatically verified as they speak with a service center executive [41]

In August 2014 GoVivace, developed speaker identification system by using voice biometric technology. This technology can be used for rapid voice sample matching with thousands or millions of voice recordings. The purpose of implementing this technology is to identify callers in enterprise contact canter settings where security is a major concern. GoVivace's SI technology is also available as an engine. Application Programming Interfaces (APIs) to use the software as a service [58]. In the UK, HSBC is developing voice recognition and touch ID services for 15 million customers by the summer in a big step towards biometric banking [42].

The popularity of voice biometric has risen more in past few years. According to Opus research, more than a half billion voiceprint will be in record, alone by 2020. People found more comfortable with biometric authentication [3].

V. CONCLUSION

This paper provides concise definition and discussion about speaker recognition technology. In addition, basic concepts of automatic speaker recognition systems, modeling technique etc. has been discussed. Speaker recognition is method of designing a system for identity of an individual through their voices. Speaker recognition has a significant prospective as it is appropriate biometric technique for security. The speaker recognition task is normally achieved by acquiring speech signal, feature extraction, modeling speech features for speaker, pattern matching and obtaining match score.

REFERENCES

[1] B. S. Atal, "Automatic Recognition of Speakers from their Voices", Proc. IEEE, vol. 64(4), pp. 460-475, 1976.

[2] Q. Jin, "Robust Speaker Recognition", PhD Thesis, Language Technologies Institute School of Computer Science Carnegie Mellon University, Pittsburgh, pp. 23-177, 2007.

[3] A. Rajsekhar G., "Real Time Speaker Recognition using MFCC and VQ", PhD Thesis, Department of Electronics & Communication Engineering, National Institute of Technology Rourkela, pp. 9-71, 2008.

[4] J. Luettin, "Visual Speech and Speaker Recognition", PhD Thesis, Department of Computer Science University of Sheffield, pp. 16-156, May 1997.

[5] T. Kinnunen, "Spectral Features for Automatic Text-Independent Speaker Recognition", Licentiate's Thesis, University of Joensuu, Department of Computer Science Joensuu, Finland, pp. pp. 1-151, December 21, 2003.

[6] M. R. Srikanth, "Speaker Verification and Keyword Spotting Systems for Forensic Applications", PhD Thesis, Department of Computer Science and Engineering Indian Institute of Technology Madras, pp. 1-135, Dec. 2013.

[7] U. Sandouk, "Speaker Recognition Speaker Diarization and Identification", PhD Thesis, University of Manchester, School of Computer Science, pp. 14-101, 2012.

[8] M. Savvides, "Introduction to Biometric Technologies and Applications", ECE & CyLab, Carnegie Mellon University. http://www.biometricscatalog.org/ biometrics/biometrics_101.pdf or biometrics_101.pdf

[9] M. El. Ayadi, Abdel-Karim S.O. Hassan, Ahmed Abdel-Naby, Omar A. Elgendy, "Text-independent Speaker Identification using Robust Statistics Estimation", Speech Communication vol-92, pp. 52–63, 2017.

[10] S. Sarkar, Sreenivasa, Rao, K., "Stochastic Feature Compensation Methods for Speaker Verification in Noisy Environments" Appl. Soft Comput. 19, pp. 198–214, 2014.

[11] G. Doddington, Liggett, W., Martin, A., Przybocki, M., and Reynolds, D, "Sheeps, Goats, Lambs and Wolves: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation", In Proc. Int. Conf. on Spoken Language Processing (ICSLP 1998) (Sydney, Australia), 1998.

[12] S. Furui, "Speaker-Dependent-Feature Extraction, Recognition and Processing Techniques", ESCA Workshop on Speaker Characterization in Speech Technology Edinburgh, Scotland, UK, pp. 10-27, June 26-28, 1990.

[13] L. P. Cordella, P. Foggia, C. Sansone, M. Vento, "A Real-Time Text-Independent Speaker Identification System", Proceedings of the ICIAP, pp. 632, 2003.

[14] B. Richard Wildermoth, "Text-Independent Speaker Recognition Using Source Based Features", PhD Thesis, Griffith University Australia, pp. 1-101, Jan. 2001.

[15] M. Breen, L. C. Dilley, J. Kraemer, and E. Gibson, "Inter-Transcriber Reliability for Two Systems of Prosodic Annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch)", Corpus linguist. ling., vol. 8 (2), pp. 277-312, 2012.

[16] A. Stolcke, "Higher-Level Features in Speaker Recognition", Winter School on Speech and Audio Processing, IIT Kanpur, Jan. 2009.

[17] N. Dehak, D. Pierre, and K. Patrick, "Modeling Prosodic Features with Joint Factor Analysis for Speaker Verification", IEEE Transactions on Audio, Speech, and Language Processing, vol. 15 (7), pp. 2095-2103, Sept. 2007.

[18] Z. Huang, L. Chen and M. Harper, "An Open Source Prosodic Feature Extraction Tool", School of Electrical and Computer Engineering Purdue University West Lafayette, pp. 2116-2121, 2006.

[19] T. Kinnunen and Haizhou Li, "An Overview of Text-Independent Speaker Recognition: From Features to Supervectors", Speech Communication vol.52, pp.12–40, 2010.

[20] N. Singh and Khan R. A., "Underlying of Text Independent Speaker Recognition", in IEEE Conference (ID: 37465) (10th INDIACom 2016 International Conference on Computing for Sustainable Global Development), held on 16th -18th March, 2016 at BVICAM, New Delhi, pp. 11-15, 2016.

[21] D. Sierra Rodriguez, "Text-Independent Speaker Identification", PhD Thesis, AGH University of Science and Technology Krakow, Faculty of Electrical Engineering, Automatics, Computer Science and Electronics, pp. 1-121, 2008.

[22] M. Ghassemian and K. Strange, "Speaker Identification - Features, Models and Robustness", Technical University of Denmark, DTU Informatics Kongens Lyngby, Denmark, pp. 1-118, 2009.

[23] D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology", MIT Lincoln Laboratory, Lexington, MA USA, @2002IEEE, pp. 4072-4075, 2002.

[24] B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and verification", Journal Acoustical Society of America, vol. 55, no. 6, pp. 1304-1312, June 1974.

[25] "Speaker Identification". *Archived from* the original *on August 15, 2014*. Retrieved September 3, 2014.

[26] A. Majetniak, "Speaker Recognition using Universal Background Model on YOHO Database", Aalborg University, The Faculties of Engineering, Science and Medicine Department of Electronic Systems, pp. 1-61, May 31, 2011.

[27] B. Martínez-Gonzalez, M. Jose Pardo, D. Julian Echeverry-Correa, Ruben San-Segundo, "Spatial Features Selection for Unsupervised Speaker Segmentation and Clustering", Expert Systems with Applications, vol. 73, pp. 27–42, 2017.

[28] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)", Speech Coding and Synthesis, pp. 495-518, 1995.

[29] P. Boersma, "Accurate Short-term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound", In Proc. the Institute of Phonetic Sciences, 1993.

[30] C. Liu, P. Jyothi, H. Tang, V. Manohar, M. Hasegawa-Johnson, and S. Khudanpur, "Adapting ASR for under-Resourced Languages using Mismatched Transcriptions", In Proc. ICASSP, 2016.

[31] E. Vayrynen, "Emotion Recognition from Speech Using Prosodic Features", PhD Thesis, University of Oulu Graduate School, University of Oulu, Faculty of Information Technology and Electrical Engineering, Department of Computer Science and Engineering, Infotech Oulu, pp. 1-92, 2014.

[32] N. Singh, A. Agrawal and Khan R. A., "Automatic Speaker Recognition: Current Approaches and Progress in Last Six Decades", Global Journal of Enterprise Information System. Vol. 9, Issue-3, July-September, pp. 38-45, ISSN: 0975-1432, 2017.

[33] D. A. Reynolds, "Automatic Speaker Recognition: Current Approaches and Future Trends" MIT Lincoln Laboratory, Lexington, MA USA, pp. 1-6, 2001.

[34] S. Memon, "Automatic Speaker Recognition: Modeling, Feature Extraction and Effects of Clinical Environment", PhD Thesis, School of Electrical and Computer Engineering Science, Engineering and Technology Portfolio RMIT University, pp. 1-242, June 2010.

[35] R. Summerfield, T. Dunstone, C. Summerfield, "Speaker Verification in a Multi-Vendor Environment", www.w3.org/2008/08/siv/Papers/Centrelink/w3c-sv_multivendor.pdf

[36] R. Nakatsu, J. Nicholson, and N. Tosa, "Emotion Recognition and its Application to Computer agents with Spontaneous Interactive Capabilities", Knowledge based systems, vol. 13, pp. 497–504, 2000.

[37] G. Gravier and G. Chollet, "Comparison of Normalization Techniques for Speaker Verification", In Proc. on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pp. 97-100, 1998.

[38] N. Singh, Khan R. A. and Raj shree, "Applications of Speaker Recognition" Science Direct, Procedia Engineering, vol-38, pp. 3122-3126, 2012.

[39] E. Shriberg and A. Stolcke, "Direct Modeling of Prosody: An Overview of Applications in Automatic Speech Processing", In International Conference on Speech Prosody, 2004.

[40] A. Michele Cavazza and C. Alberto, "Device for Speaker's Verification", http://www.google.com/patents/US4752958?hl=it&cl=en

[41] International Banking (December 27, 2013). "Voice Biometric Technology in Banking | Barclays". Wealth.barclays.com. Retrieved February 21, 2016.

[42] K. Julia, "HSBC Rolls out Voice and Touch ID Security for Bank Customers | Business", The Guardian, Retrieved February 21, 2016.