Virtual Development of Voice Analysis as a Reliable Technique: A Review

Ahuja Pooja, J. M. Vyas

Abstract— Voice Authentication technique for forensic sample is generally a challenging task for automatic, semiautomatic and human based methods. The speech samples being compared may be recorded in different situations; e.g., one sample could be a yelling over the telephone, whereas the other might be a whisper in an interview room. A speaker could be disguising his or her voice, ill, or under the influence of drugs, alcohol, or stress in one or more of the samples. The speech samples will most likely contain noise, may be very short, and may not contain enough relevant speech material for comparative purposes. Each of these variables, in addition to the known variability of speech in general, makes reliable discrimination of speakers a complicated and daunting task.

Index Terms— Voice authentication, whisper, disguising voice, speech material, reliable

1 INTRODUCTION

In 1660s the English emperor Charles I was executed in voice analysis but there was no specific and scientific reason to prove the voice authenticity.[1] Forensic researchers has developed various forensic voice analysis techniques including the effect of pressure and psychological conditions of the physical condition over voice like layered voice by criminals is a problem faced by the investigation authorities since long.[4] Now criminals started using gloves to make sure about unavailability of their fingerprints to the investigating authorities and thy make sure to burn the equipment's and traces at the crime scene to destroy DNA evidence and also disguise their voice when demanding ransom money by phone calls or in case of threatening using audio devices like Compact discs, audiotapes etc. To overcome the problems of voice identification, modern automatic methods have been developed to trace the voice phonates, frequencies and voice prints.[5] For forensic purposes a set of different speech tests can be also performed including auto and manual identification techniques and a combined result is used to make the opinion about voice
in the justice system. The main principal behind forensic voice analysis is individuality that is voice prints/ phonates is unique for every person like DNA and fingerprints. [6] Digital automated techniques uses digital wave pattern of

in university libraries, which provide good and comprehensive introductory reading. Forensic Voice Identification by Hollien, which sketches a historical background of the field and covers topics like automatic speech recognition, memory and voice lineup procedures fairly non-technical and does not require any in depth phonetic knowledge and

- *Ahuja Pooja is working as Assistant Prof. Jr. pursuing ph.d. degree program in Forensic Voice Analysis in GFSU, India, PH-9724304885 E-mail: pahuja159@gmail.com*
- *J. M. Vyas is the director general in GFSU, India, PH-07965735502. E-mail: dg@gfsu.edu.in*

Forensic Speaker Identification, considerably more technical in nature dealing with automatic speaker identification which covers some of the techniques used like cepstrum Voice analysis.[2]. Many a times there is a situation arises in court of laws regarding authenticity and scientific proof where culprit is heard but not visible in the evidence, there is an eyewitness claiming to be able to identify the culprit's voice but he is not able to guarantee, prove that the identification of voice is correct. [3] Deception and alterations in the the voice to analyse the spectra dependent analysis.

2 BACKGROUND

In late thirties, the bell telephone laboratories (BTL) resulted in an invention called the spectrograph, which was actually a result of the research, proceeded to provide aid in phonation training to the deaf and for the students learning foreign languages. In 1944 after world war-II,the term „voice print" was used for the very first time. In 1962 kersta published an article in "the nature entitled Voice print Identification" and provided opinion in court till 1967.[7] In 1967, young and kambell has challenged the research of kersta regarding accuracy of the results and revealed the results in identification of voice prints where accuracy went down up to 38%.[8] Several textbooks on forensic phonetics have been published during the last decades,many of which are found

analysis, in some depth. Statistical problems and methods involved in speaker verification evaluation like Bayesian statistics and forensic voice comparison using voice source features has gain a vide interest of the research in case of machine/instrument generated voice prints because of the raised deception and alteration in the voice by criminals using various voice stations for generating voice by making minor modifications in pitch and base of the voice.[9] Auditory phonetic approaches in forensic identification of voice have been experience based subjective analysis of voice quality. "Voice quality" is referred to laryngeal voice-tract settings and to refer to physiologically constrained as well as voluntarily. Some of the reliable features used in the pharanxial voice identification includes: Distortion features and fundamental frequency of many types i.e. absolute normalized jitter, normalized amplitude shimmer, normalized slenderness shimmer, which is referred to as the difference

in the height of approx. triangle formed due to negative spike of the glottal pulse. Singularities in the mucosal-wave consulate power spectrum of fourteen features[10]

3 STANDARD COMPARISON PROTOCOLS

The following protocols are maintained for positive outcome of voice samples and comparison of voice samples.

I. Only original recordings of voice samples can be accepted for examination, unless the original recording had been erased and a high-quality copy was still available.

II. The recordings will be played back on appropriate professional tape recorders and recorded on a professional full-track tape recorder at 7 1/2 ips. When possible, playback speed has to be adjusted to correct for original recording speed errors by analyzing the recorded telephone and AC line tones on spectrum analysis equipment. When necessary, special recorders are used to allow proper playback of original recordings that is having incorrect track placement or azimuth misalignment. Spectrograms for Voice Identification, must have standard settings and a linear expand frequency range (0-4000 Hz),wide band filter (300 Hz) and have bar display mode at least, higher specification can also be used. All spectrograms for each separate comparison should be prepared on the same spectrograph. The spectrograms must be phonetically marked below each voice sound.

III. When requires, prepare enhanced tape copies from the original recordings using equalizers, notch filters, and digital adaptive predictive deconvolution programs to reduce extraneous noise and correct telephone and recording channel effects. Prepare A second set of spectrograms from the enhanced copies and use it together with the unprocessed spectrograms for comparison.

IV. Compare similarly pronounced words between two voice samples, with most known voice samples being verbatim with the unknown voice recording. Normally, 20 or more different words are needed for a meaningful comparison. Less than 20 words usually results in a less conclusive opinion, such as possibly instead of probably.

V. The examiners have to made spectral pattern comparison between the two voice samples by comparing beginning, mean and end formant frequency, formant shaping, pitch, timing, etc., of each individual word. When available, compare similarly pronounced words within each sample to insure voice sample's consistency. Words with spectral patterns that are distorted, masked ,byextraneous sounds, too faint, or lacked adequate identifying characteristics should be eliminated.

VI. Make an aural examination of each voice sample to determine if pattern similarities or dissimilarities noted are the product of pronunciation differences, voice disguise, obvious drug or alcohol use, altered psychological state, electronic manipulation, etc.

VII. An aural comparison is taken by repeatedly playing two voice samples simultaneously on separate tape recorders, and electronically switching back and forth between the samples while listening on high-quality headphones. When a sample has a wider frequency response than the other, bandpass filters are advised to be used to compensate at least some of the aural listening tests.

VIII. The examiner should resolve any differences found between the aural and spectral results, usually by repeating all or some of the comparison steps.

IX. If the examiner found the samples to be very similar (identification) or very dissimilar(elimination), always conduct an independent evaluation by atleast one, but usually two other examiners to confirm the results. If differences of opinions still present between the examiners, additional comparisons to be done to resolve this elimination.[11] Cross-linguistic phonetic studies have yielded several insights into the possible states of the glottis.[12] People can control the glottis so that they produce speech sounds with not only regular voicing vibrations at a range of different pitches, but also harsh, soft, creaky, breathy and a variety of other phonation types. These are controllable variations in the actions of the glottis, not just personal idiosyncratic possibilities or involuntary pathological actions. What appears to be an uncontrollable pathological voice quality for one person might be a necessary part of the set of phonological contrasts for someone else. For example, some American English speakers may have a very breathy voice that is considered to be pathological, while Gujarati speakers need a similar voice quality to distinguish the word /baª|/ meaning „outside" from the word /ba|/ meaning „twelve".[13.14] Likewise, an American English

speaker may have a very creaky voice quality similar to the one employed by speakers of Jalapa Mazatec to distinguish the word /ja0!/ meaning „he wears" from the word /ja!/ meaning „tree".[15] As was noted some time ago, one Person's voice disorder might be another person's phoneme.[16] Another point on the phonation continuum exploited by certain languages (far fewer in number than languages which have voiceless sounds) is breathy voice. Breathy phonation is associated with a decrease in overall acoustic intensity in many languages, e.g. Gujarati (Fischer-Jørgensen 1967), Kui and Chong (Thongkum 1988), Tsonga (Traill and Jackson 1988), Hupa (Gordon 1998).

4 . PRACTICAL PROBLEMS WITH VOICE SAMPLES

Factors that may influence identification accuracy are primarily sample duration and acoustic quality. If we first consider the influence of sample duration, we may observe that in real life investigations samples may be very short, often just a few words or a phrase or two which means that sample duration is on the order of a few seconds. In an early study by Pollacketal. (1954), the authors observed that identification accuracy increased as sample size (for monosyllabic words) increased, but only up to about 1.2 seconds. For longer samples they claim that phonetic variation takes over as the most important factor. They conclude that "we believe that the duration of the speech sample per se is relatively unimportant, except in so far as it admits a larger or smaller statistical sampling of the speaker's speech repertoire". This is somewhat surprising finding has, however, been confirmed in other studies. In a study by Compton (1963), 15 recorded segments of the vowel [17] for each of 9 speakers, familiar to the listeners, were presented. The segments differed only in duration (25–2500 ms). For segments longer than about 75 ms, there was no increase in recognition rate as a function of duration.Bricker and Pruzansky (1966) presented stimuli which varied in duration as well as phonemic variation. They found that identification rate increased with duration only if the longer stimuli also contained more phonemic variation and that "Identification accuracy improved directly with the number of phonemes in the sample even when duration was controlled". In a study by

Orchard and Yarmey (1995) correct identification rate was substantially higher for 8 minute stimuli compared with 30 second stimuli. No attempt was made, however, to estimate the respective contributions of duration and phonological variation, but it is likely that phonological variation must have been higher in the longer stimuli. A large proportion of threats are done over the telephone and criminals often use telephones when they plan or coordinate crimes. Telephone quality speech has therefore received attention in forensic phonetics studies. Telephone lines have limited bandwidth. Most of the frequencies relevant for speech transmission are covered, but not all. Frequencies below 300 Hz are filtered out for example. With mobile phones, problems related to speech coding are introduced. These effects are particularly noticeable for female voices. Important questions in the forensic context are whether the poorer sound quality of recorded telephone conversations adversely affects voice identification and if so to what extent and how. Also, from a methodological point of view one would like to know whether one should only use voices recorded over the telephone in lineups where the incriminating call is recorded over the telephone.[18]There are surprisingly few studies that address this question, but there are some results which indicate that the problem might not be as serious as one might expect. For example Rathborn, Bull and Clifford (1981, cited in Yarmey, 1991) "failed to find any significant differences in voice identification of a target voice heard originally over the telephone and tested using a taped lineup over the telephone, in contrast to voice identification heard originally over the telephone and tested directly with a taped lineup. A question that has received some attention lately is the influence of the band-pass filtering that occurs in telephone transmissions on acoustic analysis of voice samples. In a recent study, Künzel (2001) found that the relatively high (300 Hz) lower cut-off frequency had the effect of shifting F1 in German vowels upwards compared to the corresponding tokens in a simultaneous DAT-recording. The average size of the shift was 6.6% for male and 6.1´% for female speakers and all the differences were significant at the 5% level or better. Other, but minor, artefacts were observed as well. As a consequence, Künzel warns against using formant data for speaker identification purposes if the recordings were made from telephones. His results have not been questioned, but his total rejection of the use of formant data in speaker identification based on telephone recordings has been challenged by Nolan (2002).[19]

4.1 Disguised Voice

Disguised voice up to the extent used, is a serious problem for speaker identification. In the extreme end of the spectrum we find electronic manipulation or even communicating via speech synthesis, which would make speaker identification virtually impossible. In the world of actual forensic work, however, voice disguise tends to be of a rather

unsophisticated nature. Künzel (2000) notes, based on experience from BKA (the German Federal Police Office), revealed that "falsetto, pertinent creaky voice, whispering, faking a foreign accent, and pinching One's nose" are the most common types. Basically the same observations have been made in experimental studies. In a study by Masthoff (1996) where undergraduate students served as subjects, the majority of the chosen disguises (35%) were phonation level disguises (whisper, raised pitch or lowered pitch). Articulation level disguises (dialect mimicry, foreign accent etc.) were also used (20%). The remaining disguises were combinations of two types. Electronically manipulated messages are still rare, but Künzelnotes that there has been an increase in recent years, mainly in the form of editing recorded voices. Even if the used types of disguise in most cases are rather unsophisticated, disguise may nevertheless have a considerable detrimental effect on speaker identification. In a study by Reich and Duke (1979) where various types of disguise were used, all types produced significantly less correct identification. Hyper nasality produced the greatest effect but there were in most cases no significant differences between the different types. Whisper, one of the more common types, resulted in markedly less correct identification in a study by Orchardand Yarmey (1995) if whispered samples were compared with phonated samples. If both the reference and the test samples were whispered the difference was less pronounced. Voice disguise is not as common as one might think. Künzel (2000) reports that:" Over the last two decades, between 15 and 25 per cent of the annual cases dealt with at the BKA speaker identification section exhibited at least one kind of disguise".[20] Voice identification by manual methods has shown variability in result accuracy based on the examiners experiences and skills. Automatic and spectrographic identification techniques have been introduced in the identification, where a sound spectrograph is used for identification of voice which produces a visual graph (voice spectrogram) of the speech as a function of time on horizontal axis and frequency at vertical axis having voice energy in grey scale/colour differences.[21]it is a well-accepted research tool in voice identification i.e. used to study individual vowel characteristics, physiological speech anomalies etc. the spectrographic voice identification assumes that intra-speaker variability including differences in the same utterance repeated by the same speaker is discoverable from inter-speaker variability of the differences in the same utterance by different speakers.[22]

4.2 Tilt in Voice Spectra

One of the major acoustic parameters that reliably differentiate phonation types in many languages is spectral tilt, i.e. the degree to which intensity drops off as frequency increases. Spectral tilt can be quantified by comparing the amplitude of the fundamental to that of higher frequency harmonics, e.g. the second harmonic, the harmonic closest to the first formant, or the harmonic closest to the second formant. Spectral tilt is characteristically most steeply positive for creaky vowels and most steeply negative for breathy vowels. In other words, the falloff in energy at higher frequencies is least for creaky voice and most for breathy voice. Subtracting the amplitude of the fundamental from the amplitude of higher harmonics thus yields the greatest values for creaky vowels and the smallest values for breathy vowels, with intermediate values for modal vowels.Spectral tilt reliably differentiates phonation types in a number of languages, including Jalapa Mazatec (Kirk et al. 1993, Silverman et al. 1995), which contrasts creaky, breathy, and modal vowels, (Bickley 1982, Ladefoged 1983, Jackson et al. 1988), which distinguishes between breathy and modal vowels (as well as a third type of phonation, strident,), Gujarati (Fischer-Jorgensen 1967), which contrasts breathy and modal vowels, Kedang (Samely 1991), which contrasts modal and breathy vowels, Hmong (Huffman 1987), which distinguishes breathy and modal vowels, Tsonga (Traill and Jackson 1988), which contrasts breathy and modal nasals, some minority languages of China (Jingpho, Haoni, Wa, Yi) examined by Maddieson and Ladefoged (1985), which contrast a "tense" phonation somewhat different from creaky phonation with a more modal phonation type, and, finally, Mpi, which also contrasts tense and non-tense (or "lax") phonation. Different measures of spectral tilt do not always behave uniformly in differentiating phonation types in a single language. In Mpi, which uses tone contrastively, Blankenship (1997) found interactions between tone level and measurements of spectral tilt. The amplitude difference between the fundamental and the second harmonic was a more reliable indicator of phonation type for high tone than for either mid or low tone, whereas the amplitude difference between the fundamental and the harmonic closest to the second formant was more useful for differentiating phonation contrasts in mid and low tone vowels than in high tone vowels. Investigation of phonation differences is an important area of research, as many languages employ distinctions which rely solely on differences in voice quality. As we have seen, these distinctions may involve two or more different phonation types and may affect consonants, vowels, or both consonants and vowels. In addition, many other languages regularly use non-modal phonation types as variants of modal voice in certain prosodic contexts. Languages also differ in their timing of non-modal phonation relative to other articulatory events in interesting ways, although there are certain recurrent timing patterns

and distributional restrictions which warrant explanation. Differences in phonation type can be signaled by a large number of quantifiable phonetic properties in the acoustic, aerodynamic, and articulatory domains, the last of which has been relatively unstudied due to the invasive measurement techniques required. It is unlikely; however, that future research will yield many truly universal observations about the range and realization of phonation types in languages of the world. We can never know whether some language in the past had or in the future will have a novel method of using the vocal folds to make a linguistic contrast. The occurrence of phonetic rarities such as the strident voice quality and few neighboring languages shows that we can use the glottis in totally unexpected ways.[23,24,25]

5 Conclusion

Voice being a part of behavioral biometrics is virtually developing Forensic importance in cases of extortion, bribery etc. cases. Voice is used as corroborative evidence is considerable trustworthy source of evidence. Forensic Voice analysis of today is based on overall outcome based on principals and experiments of scientific modulus. As per the studies related to the subject, it extents that limitations and errors are major issues for voice identification. Positive and prominent results are identifiable if ideal voice samples with enough speech length and vowel counts are obtained. Demonstration of frequency vs. time spectra's of words and vowel/consonants makes it a more reliable technique.

Acknowledgment

I am grateful to staff of Directorate of Forensic Science, Gandhinagar, Gujarat for their technical support. I extend my thank to the director, IFS M. S. Dahiya for his guidance and enthusiasm

References

[1] . Alexander, A., Botti, F., and Drygajlo, A. (2004). Handling Mismatch in Corpus-Based Forensic Speaker Recognition. In Odyssey 2004, The Speaker and Language Recognition Workshop, pages 69–74, Toledo, Spain.

[2] . Arcienega, M. and Drygajlo, A. (2003). A Bayesian network approach for combining pitch and reliable spectral envelope features for robust speaker verification. In Kittler, J. and Nixon, M. S., editors, Proc. 4th Int. Conf. on Audio- and Video- Based Biometric Person Authentication, pages 78–85, Guildford, UK. Springer.

[3] . Dunn, R. B., Quatieri, T. F., Reynolds, D. A., and Campbell, J. (2001). Speaker recognition from coded speech in matched and mismatched conditions. In 2001: A Speaker Odyssey, Crete, Greece.

[4] . K¨unzel, H. J. (1998). Forensic speaker identification: A view from the crime lab. In Proceedings of the COST Workshop on Speaker Recognition by Man and Machine, pages 4–8, Technical University of Ankara, Ankara, Turkey.

[5] . Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits in our Capacity for Processing Information. The Psychological Review, 63:81–97

[6] . Oglesby, J. Mason, J. (1989). Speaker recognition with a neural classifier. In Proceedings First IEE International Conference on artificial Neural Networks, volume 313, pages 306–309.

[7] . Loevinger, L. (1995). Science as evidence. Jurimetrics, 35(2):153–190.

[8] . Martin, R. (1994). Spectral subtraction based on minimum statistics. In EUSIPCO- 94, pages 1182–1185.

[9] . Rossy, Q. (2003). Simulation de cas reels de reconnaissance de locuteurs au moyen du logiciel ASPIC. Project report, Institut de Police Scientifique, Ecole des Sciences Criminelles, University of Lausanne, Switzerland.

[10] . Zimmermann, P. (2005). Analyse de l"influence des conditions d"enregistrement dans la reconnaissance automatique de locuteurs en sciences forensiques. Project report, Institute Police Scientifique, Ecole des Sciences Criminelles, University of Lausanne, Switzerland.

[11] . Meuwly, D. (2000). Voice Analysis, in : Encyclopedia of Forensic Science, pages 1413 – 1420. London: Academic Press Ltd.

[12] . Titze, I.R. (1994). Principles of Voice Production, Prentice Hall (currently published by NCVS.org), ISBN 978-0-13-717893-3.

[13] . Sundberg, Johan, The Acoustics of the Singing Voice, Scientific American Mar 77, p82

[14] . Greene, Margaret; Lesley Mathieson (2001). The Voice and its Disorders. John Wiley & Sons; 6th Edition. ISBN 978-1-86156-196-1.

[15] . Rothenberg, M. The Breath-Stream Dynamics of Simple-Released Plosive Production, Vol. 6, Bibliotheca Phonetica, Karger, Basel, 1968.

[16] Titze, I. R. (2006).The Myoelatic Aerodynamic Theory of Phonation, Iowa City:National Center for Voice and Speech, 2006.

[17] . Phil Manchester (January 2010). "An Introduction To Forensic Audio". Sound on Sound.

[18] . Maher, Robert C. (March 2009). "Audio forensic examination: authenticity, enhancement, and interpretation". IEEE Signal Processing Magazine 26: 84–94.

[19] Alexander Gelfand (10 October 2007). "Audio Forensics Experts Reveal (Some) Secrets". Wired Magazine.

[20] . Labov, William (1972) Sociolinguistic patterns. Philadelphia, PA: University of Pennsylvania Press, p192.

[21] . Eagleson, Robert. (1994). 'Forensic analysis of personal written texts: a case study', John Gibbons (ed.), Language and the Law, London: Longman, 362–373.

[22] . Gibbons, J., V Prakasam, K V Tirumalesh, and H Nagarajan (Eds) (2004). Language in the Law. New Delhi: Orient Longman. Koenig, B.J. (1986) 'Spectrographic voice identification: a forensic survey', letter to the editor of J. Acoustic Soc, Am., 79, 6, 2088- 90.

[23] Koenig, B.J. (1986) 'Spectrographic voice identification: a forensic survey', letter to the editor of J. Acoustic Soc, Am., 79, 6, 2088-90.

[24] . Pennycook, A. (1996) 'Borrowing others words: text, ownership, memory and plagiarism', TESOL Quarterly, 30, 201-30.

[25] . John Olsson (2004). An Introduction to Language Crime and the Law. London: Continuum International Publishing Group.