Lecture notes- Approaches to speaker recognition

Various methodologies for approaching the problem of speaker identification have been proposed. For identification purpose, different well recognised standard techniques will be used for maintaining the validity of the work done and the choice will be as per the requirement:

- ***Listener method or auditory analysis-***

  The voice of a person is as easily distinguishable by the ear, as face by the eye. This method of speaker recognition by listening is the oldest amongst all. In this situation a person attempts to recognize a voice by its familiarity. This ability of humans to recognize many familiar people by their voices is exceptional both in accuracy and adaptability. In this method, the judgement regarding the similarity and dissimilarity between the two speech events is taken after hearing the samples again and again by the experts to find out some similarities in their linguistic, phonetic and acoustic features. The different utterances of the speakers are segregated in respect of each speaker by way of repeated listening of recorded conversation. The segregated conversations of each speaker are repeatedly heard to identify linguistic features and phonetic features like articulation rate, flow of speech, degree of vowels and consonant formation, rhythm, striking time, pauses etc. There are cues in voice and speech behaviour, which are individual and thus make it possible to recognize the familiar voices. These clue words are selected from both questioned and specimen samples of the speaker and are then used for instrumental analysis.

  Human listeners are robust speaker recognizers when presented with the degraded speech. Listener performance free from all types of limitations like the signal to noise ratio, speech bandwidth, the amount of speech material, distortions occurring in the speech signals as a result of speech coding, transmission systems, etc. This is owing to the fact that there are some knowledge sources that can support this speaker recognition system in many ways; providing weak, moderate and high discriminating power. Auditory speaker recognition has long been used and accepted in forensics as part of the testimony of a victim or witness. Prior to the inventions of the telephone and sound recording equipment, it could be the key evidence on behalf of which a suspected individual could be identified or excluded from an offence

committed in the dark or when a victim has been blindfolded. However, with any human decision process, it is generally believed that the auditory analysis by a listener leads to a subjective decision. Even after that, this method is still being used in various countries for the purpose of identifying a criminal forensically. In this method, the identification is done on the basis of following voice characteristics-

- *Quality of speech sample*- Synthetic speech can be compared and evaluated with respect to intelligibility, naturalness, and suitability for used application. The quality of a person's voice depends upon the pronunciation, Rhythm, Grammar, Speech sounds like vowels and consonants, plosives, fricatives, nasal sounds, nasalized sound, accent, fluency, intonation, Phrasing and Blending. Each person possesses a unique voice quality which depend on number of anatomical features, such as, dimension of oral tract, pharynx, nasal cavity, shape and size of tongue and lips, position of teeth, tissue density etc.
- *Linguistic features*- Linguistics is the scientific study of natural language. These features involves, the stylish impression of speech, delivery of speech (the style in which the speech is delivered i.e., Manuscript, Memorized, Impromptu, and Extemporaneous), Phonation (the process by which the vocal folds produce certain sounds through quasi-periodic vibration or any oscillatory state of any part of larynx that modifies the airstream, of which voicing is one example).
- *Articulatory speech*- This is a type of speech produced by movement or articulation of the articulators. This involves, flow of speech (depends upon the fluency of the speaker), plosive formation (First, a complete closure of the passage of air at the same point in the vocal tract, then the removal of the closure, causing a sudden release of the blocked air with some explosive noise), nasality (Nasal consonants have a continuous full closure at some point in the oral cavity. Since the velum is set in the low position, opening the velopharyngeal port, air is let out through the nasal cavity).
- *Prosodic analysis*- It involves the intonation pattern, dynamic of loudness (dynamics refers to the volume of a sound or note and loudness is the strength of sensation received through the ear), speech rate (relative timing of different speech events in spoken utterances), speech variations, striking time features, pauses (number/length/pattern).
- *Voice impairment*- Speech or language impairment (SLI) referred as inability of a speaker to communicate properly that adversely affects a person's educational

performance. Some of the important causes of such type of disorders include loss of senses that perceives sound, effect of mental state, damage to brain, neurological disorders, influence of drug , and vocal abuse or misuse.

- *Temporal measurements*- The temporal properties of speech play an important role in linguistic contrast. Speech is comprised of three main temporal features including envelope, periodicity and fine structure. Each feature has distinct acoustic manifestations, auditory and perceptual correlates and roles in linguistic contrasts. These measurements involves phonation-time (P/T) ratio, speech time (S/T) rate, speech burst (its number/length/patterns).

Experts working in several government forensic laboratories including laboratories in Germany, Austria, Sweden, the Netherlands and Spain, and in private practice in countries like the United Kingdom and Germany, are still practising this phonetic-acoustic technique for identification.

- ***Instrumental analysis or spectrographic method-***

The spectrographic method for speaker recognition makes use of an instrument that converts the speech signals into a visual display. The identity of a speaker is established on the combined effort of aural and spectrographic analyses**.** In 1941, an electro mechanical acoustic spectrograph was developed by Dr. Raleph Potter, Bell Telephone Laboratory, with an idea to convert sounds into pictures.

A sound spectrograph is an instrument which is able to give a permanent record of changing energy-frequency distribution throughout the time of a speech wave. The spectrograms are the graphic displays of the amplitude as a function of both frequency and time. They convey information about the message by the speaker as well as about the speaker himself. The spectrograph is more commonly known as the Voiceprint analyser. Voice patterns are transformed into visual patterns on a graph that moves through an instrument at a controlled speed, and patterns drawn on the paper as it moves. By analysing the charts, you can compare a tape of an individual's normal speech pattern with a tape of the same person being questioned about his or her involvement in some type of crime or other misbehaviour. These voiceprints may be an important in helping the law enforcement agencies in identifying the criminals. Much like fingerprints, voiceprint identification uses the unique features in the spectrographic impressions of people's utterances.

In the classical analogue spectrograph a magnetic tape recorder and playback unit is used to process the sounds into electronic signals. These signals are then sent through a variable electronic bandpass filter, which selects a frequency band that is to be analysed, before a stylus measures its energy and records the results on electrical sensitive paper. The paper is mounted on a drum, which is rotating during playback in order to plot the time variations in the signal. When the whole length of the speech sample in analysed at a specific frequency band, the band of the filter and the position of the stylus are correspondingly altered. The tape is then played again in order to analyse a new part of the frequency spectrum. This process is repeated over again until the entire desired frequency range is analysed. In each spectrogram, the horizontal dimension is time, the vertical dimension represents frequency and the darkness represents the intensity on the compression scale. The differences in amplitude values are shown in a grey scaling where black represents the most intense and white the least intense waveform components.

However since 1962, it was considered as a fool- proof method of personal identification, voice identification by spectrographic analysis, the "voiceprint" technique has been in a legal limbo. But the recent developments in both science and the law, however, indicate that despite initially adverse scientific and judicial reaction, spectrographic voice identification is perhaps coming of legal age.

In this method, a trained examiner may be able to give an opinion about the similarity between the two samples on the basis of characteristics like:

- *Fundamental frequency*- It is the frequency of vibration of vocal cord produced during the rapid opening and closing of vocal cord. The fundamental frequency of a periodic signal is an inverse of period length. The period, in turn, is the smallest repeating unit of a signal. In voice spectrogram, horizontal distance between vertical striations is an indication of fundamental frequency. It also includes the pitch of voice i.e., the vocal cords vibration rate.
- *Formant frequency*- Formants are the spectral peaks of the sound spectrum. Formant also used to mean as an acoustic resonance and in speech science and phonetics, a resonance of human vocal tract. A formant represents the amount of acoustic energy at a particular frequency in the speech wave. Formants can be seen very clearly in a wideband spectrogram, where they are displayed as dark bands. The darkness of the formant represents the effectiveness of the spectrogram.

Formant frequency is the average frequency of the formant measured generally at the middle of the formant band.

- *Amplitude*- The amplitude of a sinusoidal waveform is the vertical distance from zero to maximum (peak) displacement. In sound waves this is related to what we perceive as volume, and its equivalent is the magnitude of pressure. It is the objective measurement of the degree of change (positive or negative) in atmospheric pressure (the compression and rarefaction of air molecules) caused by sound waves. These atmospheric pressure variations i.e. from high pressure to low pressure can be successfully observed in the sounds with greater amplitude. Amplitude is the perfect representation of acoustic energy or intensity of a sound.
- *Intensity*- Degree of darkness in the spectrogram reflects the sound energy at a particular region. The amount of sound energy produced is dependent on the amplitude and the frequency, which can be explained by the formula:

$$I^2 = A^2 * F^2$$

Where, I is the Intensity i.e., energy of the sound wave

A is the Amplitude of the sound wave

F is the Frequency of the sound wave

- *Loudness*- It is the psychological correlate of intensity. It is usually considered as the strength of sensation received through the ear. In general, two tones, one having the greater intensity will sound louder, if the pitches are not too far apart.
- *Bandwidth*- It is the difference between the highest formant frequency and the lowest frequency measured at the formant band.
- *Transitional characteristics*- These are the rapid change, especially of the second formant frequency that occurs between certain phonemes.
- *Tones and noises*- Tones are made up of sound that has periodic vibrations. The best example of this type of sound is the vibration made by a tuning fork. For low tones, the pitch decreases with intensity but for high tones the pitch increases with intensity.

  The sounds, does not have the regularity of vibration, which has no periodicity and no identifiable pitch, is usually classified as noise. E.g.: - the crash of a falling tree.
- *Spectral correlation*- Degree of correlation between short time spectra at different frequencies has also been found to be speaker sensitive.

It was observed that the use of spectrographic pattern matching and aural comparison in forensic voice identification requires careful control of examiner bias and an awareness of the principles of signal detection theory.

However, the main drawback of this voiceprint analysis is that the spectrograms of the speech signal from same individual will show large intraspeaker variations, because of the fact that no speaker actually is capable of producing two identical speech utterances. This method is obviously neither objective nor superior to aural-perceptual methods; it is basically a shifting of subjective judgement to the visual domain. The objectivity, reliability and validity of the method have been discussed controversially. The method has been widely used in the US, parts of Europe and other countries until the 1980s but in the present scenario it has been losing its ground. The FBI are using it for investigative purposes, most U.S. courts do not accept voiceprint evidence. Today voiceprint identification is not used in forensic labs in Europe, but still practised in developing countries like China, Vietnam etc.

Spectrographic Evidence

Most of the information available about the usage of spectrographic voice identification evidence in the court of law has its origin in different jurisdictions in the USA. Reasons for this include the fact that the technique was developed and has been most widely used in the USA, but also that most of the leading scientists and studies performed regarding the subject have been located there. Spectrographic evidence was first found admissible in 1966 in the case of People v. Straehle. The expert witness in this case was Kersta who testified on the basis of his own voiceprinting technique. This was the start of a constant controversy regarding the admissibility of spectrographic evidence that still exists today. Based on different case conditions and interpretations of the governing tests of scientific evidence, different jurisdictions have either ruled for or against admissibility. The arguments opposing admissibility have mostly lain on the principle of general acceptance and the fact that the possible persuasive force of the evidence does not correspond to its reliability.

General Acceptance

The principle of general acceptance of scientific evidence, as stated in the Frye test, is based on “a well recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs”. When this test has been used on

the admissibility of spectrographic evidence, there have been differences in its interpretations that have lead to different decisions. Some courts have ruled against admissibility on a basis that general acceptance is required from scientists with a broad theoretical knowledge. With this interpretation of the field in which spectrographic voice identification belongs, a uniform acceptance would be required from scientists in physiology, phonetics, linguistics, psychology, medicine and engineering. Other courts have found spectrographic evidence admissible on a basis that the general acceptance is only required from those who are familiar with the usage of the technique. This might seem a more reasonable approach, but does however impose a problem, as there exist today only a limited amount of practitioners and scientists that would be qualified to form an opinion. There have also been differences in the interpretation of the term general acceptance. Some courts have rejected spectrographic evidence because there has not been a uniform opinion on its reliability in the relevant scientific community, while others have admitted the evidence even when presented with substantial scientific disagreements. There is another issue of general acceptance worth considering, although not generally confronted by the courts of law: Does the general acceptance criteria apply to the technique itself, or should it also include the underlying scientific principle? Voice identification in general, as mentioned earlier, is performed on the assumption that intraspeaker variability is less prominent or different to intraspeaker variations. Since this is a theory that has not been sufficiently scientifically validated, a general acceptance of the underlying principle of voice identification cannot exist.

Persuasive Force

Another important concern about the admissibility of spectrographic evidence is the possibility of overvaluation of its conclusiveness by the fact finder. Because of the similarities to fingerprint identifications and the technical complexity of the technique, it is believed that a lay jury might find it difficult to assess its strength and weaknesses. It would normally be the job of the opposing attorney to deal with the persuasive force of experimental evidence, through cross-examination and presentation of opposing expert testimonies to the court. Since this is not always possible, other alternative approaches have been conducted in some cases where spectrographic evidence has been found admissible. In some cases instructions have been given to the jury on how to deal with the evidence. Dependent on the particulars case, the contents of these instructions have normally been that the jury can choose to disregard the evidence as

opinion only, either on the basis of uncertainties about accuracy or the examiners lack of scientific qualifications. Another approach has been to admit the spectrographic evidence only in cases where it is collaborating with other evidence. If this is used to correct for uncertainties regarding the accuracy of the technique, it is crucial that the examiner has no prior knowledge about the case, before making the comparisons and conclusion.