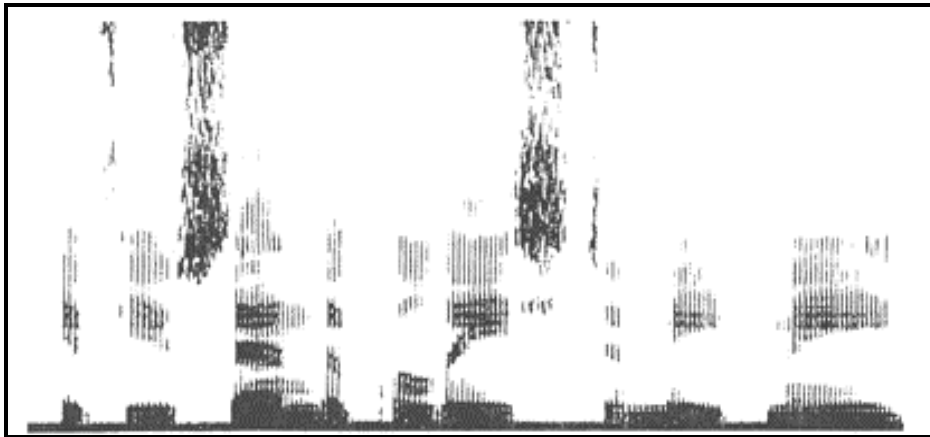


# **Forensic Voice Identification**



A spectrogram of the phrase "If you send the police to meet me..."

**A Research Paper in Forensic Science**  
**The University of Auckland, New Zealand**  
**2000**

*By Yme Asgeir Kvistedal*

Supervised by Dr. Douglas Elliot

# **Table of Contents**

<b>Introduction</b>	<b>5</b>
 <i>Part 1:</i>	
<b>The Vocal Organs</b>	<b>7</b>
The Respiratory System	8
The Larynx	8
The Pharynx	9
The Nasal Cavity	9
The Oral Cavity	9
<b>Phonation</b>	<b>10</b>
Voiced Sounds	10
Unvoiced Sounds	10
Alternative modes	11
<b>Articulation</b>	<b>12</b>
Vowels	12
Diphthongs	13
Consonants	14
Plosives	14
Nasals	14
Fricatives	14
Approximants	15
<b>The Nature of Sounds</b>	<b>16</b>
Frequency and Waveforms	16
Amplitude and Intensity	16
Filters	17
Resonance	18
<b>Acoustics of Speech Sounds</b>	<b>19</b>
Formant Structures	19
Formant Transitions	20
Noised Sounds	21

## ***Part 2:***

<b>The Theory of Voice Identification</b>	<b>22</b>
Interspeaker Variations	22
Intraspeaker Variations	23
Scientific Basis	23
<b>The Sound Spectrograph</b>	<b>24</b>
Analogue Systems	24
Accuracy and Resolution	24
Filter Modes	25
Digital Systems	25
<b>Spectrographic Identification Studies</b>	<b>27</b>
Kersta's Voiceprinting Method	27
Michigan State University Voice Identification Project	28
Voice Disguises	29
The Classification and Filing System	29
Forensic Identifications	30
National Academy of Science	31
FBI's Performance Survey	31
Summary	31
<b>Voice Comparison Standards</b>	<b>33</b>
Examiner qualification	33
Reference Samples	33
Exclusion of Samples	34
Aural Comparisons	34
Preparation of Spectrograms	35
Spectrographic Comparisons	35
Conclusions	36
Testimony	36
<b>Spectrographic Evidence</b>	<b>37</b>
General Acceptance	37
Persuasive Force	38

### ***Part 3:***

<b>Developments in Voice Identification</b>	<b>39</b>
Digital Sound	39
Spectrum Analysis	40
Parameter Selection	40
Parameter Performance	41
<b>Statistics of Voice Identification Evidence</b>	<b>42</b>
Data Base	42
Error Rates	42
Likelihood Ratios	43
<b>Conclusion</b>	<b>45</b>
<b><i>Appendix A:</i></b>	
Phonemes in Standard English	46
<b><i>Appendix B:</i></b>	
Reference Articles	47
Reference Books	49

### **Table of Illustrations**

Front page: A spectrogram of the phrase (Miles 1989)	1
Figure 2: Diagram of the vocal tract (Nature 1962)	7
Figure 3: The larynx (Clark and Yallop 1995)	10
Figure 4: Vocal fold settings (Ball 1993)	11
Figure 5: The vocal cavities (Clark and Yallop 1995)	12
Figure 6: The cardinal vowel diagram (Clark and Yallop 1995)	13
Figure 7: Examples of waveforms (Clark and Yallop 1995)	17
Figure 8: Resonance response (Ball 1993)	18
Figure 9: The source and filter model (Clark and Yallop 1995)	19
Figure 10: Acoustic Properties (Clark and Yallop 1995)	20
Figure 11: Three spectrograms (N.A.S. 1979)	26
Figure 12: Spectrum Analysis (Clark and Yallop 1995)	44

## **Introduction**

In a society where electronic communication and surveillance equipment are commonly used, both in criminal activity and for investigational purposes, analysis of recorded evidence has become an important part of forensic science. In forensic audio, which is the common name of these analyses, the sciences of acoustics, electronics and linguistics are combined in order to perform procedures such as tape enhancements and authentications, transcription and interpretation of conversations, gunshot acoustics and voice identification. Forensic voice identification, which is the subject of this paper, becomes an issue during investigation or litigation when there exists a recording in connection to a crime where the identity of the speakers are in question. These can include situations such as drug deals where conversations are recorded with the aid of a body wire, telephone threats that have been recorded by the recipient, robberies which have been taped by security cameras or any other crime where there exists recorded evidence containing unknown human voices.

This paper is divided into three parts, where the first part is an introduction to acoustics and speech production. This is included in order to get a good understanding of the underlying scientific principles of voice identification and the methods by which it is performed. In speech sounds, the markers used in forensic analysis will be acoustical events, which are the result of the mechanisms used in speech production. Ideally these markers ought to satisfy the demands of being: highly discriminating, consistent with the speaker, consistent through time, easy measurable, unaffected by the speakers health and unaffected by environmental noise. As will be shown, none of the markers identified to date satisfy all of these demands, and the quality of the markers are highly dependent upon what mechanisms are involved in their production. In transcribing speech sound it is common to use phonemes from the universal phonetic alphabet. Depending upon dialect there are differences in how English words are transcribed. In this paper the phonemes and transcription of standard English have been used, and a complete table with examples of these are enclosed in appendix A.

The second part of the paper is a review of the traditional method of voice identification performed in forensic science. This involves a comparison between the unknown voice and reference samples of a known voice by both aural and graphical means. The graphical plots used in these comparisons are often referred to as voiceprints, which is a name given because of the alleged similarity to fingerprints. The voiceprints are, in this paper, referred to by their original name: spectrograms. This because the term voiceprint is misleading with regards to accuracy, as spectrographic voice identification is by no means as conclusive as fingerprinting. Some of the procedures that are often part of a spectrographic voice comparison are not discussed in detail in this paper. These include tape enhancements, quality and authentication analysis. These are excluded because they do not have any direct influence on the issue of reliability. In addressing the unsettled issue of admissibility of spectrographic evidence in the court of law, there will be no description of specific cases where this issue has been raised. However, some of the most

frequently encountered arguments both for and against admissibility, and the ways these have been handled by different courts, will be discussed.

The third part of the paper contains a description of some of the techniques developed for automatic speaker recognition and how these could be applied in a forensic context. This is an area of speech science that has received a lot more attention from the scientific community than spectrographic voice identification. It is also anticipated that this will continue into the future as part of an attempt to make it possible for humans to communicate with computers by voice. Although some of the developed techniques have been used as the basis for expert evidence on voice identification, there has been no attempt of involving these in a standardised forensic procedure. In the third part of the paper there is an outlined of a proposed new method on how the different techniques on automatic speaker recognition could be combined in a forensic context. This involves dynamic analysis with statistical evaluations of the evidence for each case, which can result in either estimated error rate or Bayesian likelihood ratio. This approach has the potential of overcoming some of the most important concerns regarding spectrographic voice identification on the issue of admissibility. The outlined approach is based upon some assumptions that have to be tested and evaluated through further studies.

## *Part 1:*

### **The Vocal Organs**

The speech chain starts with activity in the nervous system, which sends signals to the different organs in the vocal tract involved in the physical production of speech. Air is pressed up from the lungs through the larynx, where it gets converted into sound waves by the vocal folds. It then passes through the pharynx, the oral and nasal cavities, where the waves develop their final form, before they leave as speech through the lips and the nose. A common characteristic of the organs involved is that they have other primary functions than production of speech. As these are not in the scope of this essay I will not discuss them in detail. I will also only concentrate on the organs in the vocal tract, as these are the sole ones relevant for voice identification purposes.

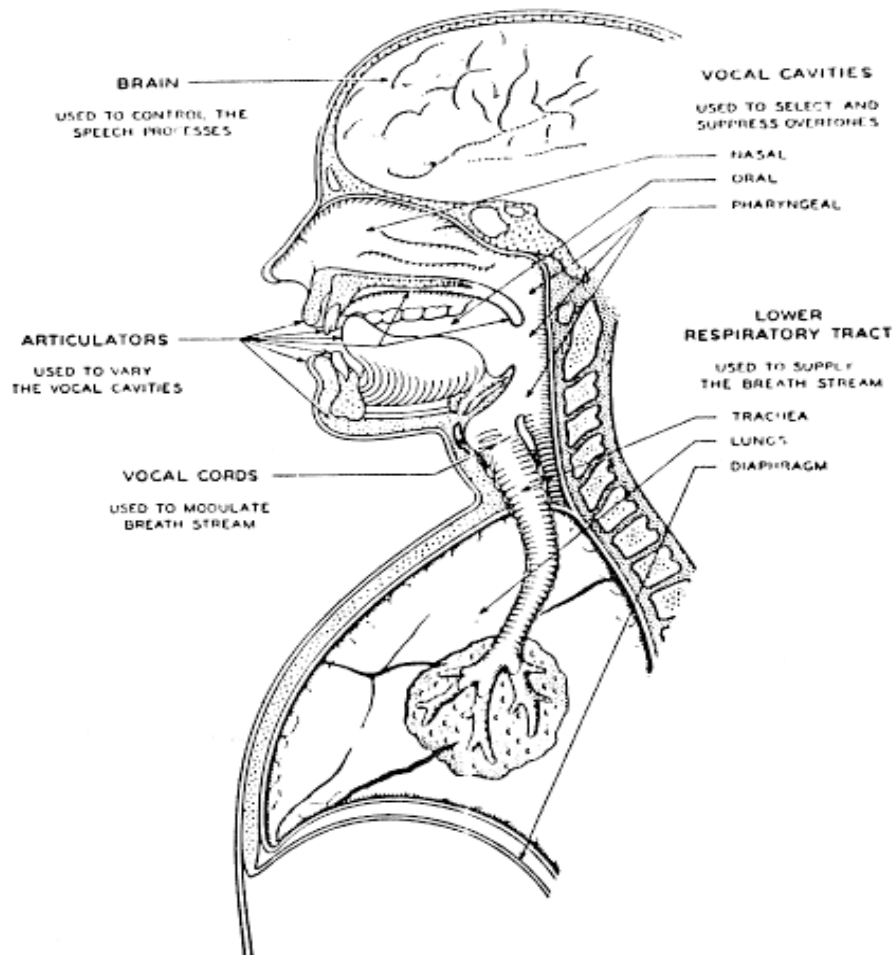


Figure 2: Diagram of the vocal tract.

## **The Respiratory System**

The respiratory system is located in the thoracic cavity and consists of the lungs and their connection to the digestion system, the trachea. Its primary function is to bring oxygen to the blood and remove the blood's contents of carbon dioxide, but in speech production it works as the main source of airflow required in order to create sounds. The two lungs are made up of many small elastic air sacs, called alveoli. These open into small tubes, which again open into larger and larger tubes before two tubes, termed the bronchi, eventually become united at the base of the trachea. The trachea is a single tube surrounded by cartilage rings. It is typically about eleven centimetres in length, and connected to the larynx at the opposite end. During the respiratory cycle, muscular forces are first applied to the thoracic cavity, in order to increase its volume. The alveoli then expand to cause a low air pressure in the lungs ( $P_{sg}$ ), compared with atmospheric pressure. To maintain equilibrium, air flows in to the respiratory system. When the muscles again retain relaxation, the volume of the thoracic cavity will decrease.  $P_{sg}$  is once again higher than atmospheric pressure, and air is expelled from the lungs. During speech  $P_{sg}$  is relatively consistent, but will vary with the speech's loudness, as an increase in intensity requires a corresponding increase in air pressure.

## **The Larynx**

The larynx is the valve of the respiratory system, which closes during eating and drinking so that food or liquid will not enter the trachea. In speech production it acts as a phonatory mechanism, transforming the airflow from the respiratory system into waveforms. The base of the larynx is a ring called the cricoid cartilage and on top of this there is a structure called the thyroid cartilage. The thyroid is often described as looking like a snowplough, the front being visible in males as the Adams apple (it is not visible in females because of the plough's larger front angle). It is connected to the cricoid by one joint on each side, which allow mobility. At the rear end of the cricoid there are two pyramid structures, called the arytenoid cartilages. These also possess mobility in the form of rotating and sliding motions. Between the arytenoids and the front part of the thyroid, there are two parallel muscles called the vocal folds (vocal cords). These are flexible, and the combined movement of the arytenoids and the thyroid can adjust their length, shape and tension. It is through the space between the vocal folds, the glottis, which the air from the respiratory system passes and gets converted into waveforms during voiced speech. Normally, tensions in the vocal folds will oppose opening of the glottis, but when  $P_{sg}$  increases they are forced apart. When air passes through this narrow opening in the glottis, its acceleration causes the pressure to drop. This results in a new closure of the glottis. As this cycle repeats itself, the vocal folds start to vibrate with a frequency determined by the combined tension, length and mass of the vocal folds and the increase in  $P_{sg}$ . Speech sounds, which are the puffs of air forced through the glottis, will have a fundamental frequency (pitch) of the same value as the frequency of the vocal fold vibrations. Loudness is also determined by the vocal folds, increasing when the glottis is kept closed for a longer period during each vibration cycle. When the vocal folds are

forced apart, the glottis will have a wider opening, resulting in higher amplitude as more air is let through.

## **The Pharynx**

The pharynx is the passage for air to the respiratory system and for food and liquid to the stomach. In speech production it acts as one of three resonating chambers that amplifies the intensity and shapes the waveforms into the different sounds, as they are known in speech. The pharynx is a tube, typically twelve centimetres long. For description purposes it is commonly divided into three parts. The inferior laryngo-pharynx and the oro-pharynx (middle section), connect the larynx to the rear end of the oral cavity. These sections can be altered in volume, as their diameters are dependent on the position of the tongue and their lengths vary with the up and down movements of the larynx. The superior naso-pharynx is the extended connection to the nasal cavity and can be sealed off from the lower sections by raising the velum.

## **The Nasal Cavity**

The nasal cavity is another resonance chamber, which reaches from the pharynx to the nostrils. It is typically about ten centimetres long, and has a complex cross sectional shape with a large surface area used to warm and humidify air during respiration. There are no muscular structures in the nasal cavity that can be used to alter its shape or size. This means that its resonating characteristics are fixed and cannot be deliberately altered, though external factors, such as flu infections can be influential. The velum is a muscular structure located at the opening between the pharynx and the nasal cavity (the velopharyngeal port) and its tip is often referred to as the uvula. As already mentioned, raising the velum can completely or partially seal off the nasal cavity from the rest of the vocal tract, creating accordingly nasal and nasalized sounds.

## **The Oral Cavity**

The oral cavity is the third resonating chamber and the single most important part of the vocal tract in production of the phonetic contents in speech. It is defined as reaching from the top of the oro-pharynx to the lips. It is separated from the nasal cavity by the roof of the mouth, which consists of the alveolar ridge, hard palate and velum. The phonetic importance of the oral cavity relies on our ability to modify its volume and shape with movement of the different articulators. The most important of these is the tongue, which is involved in the production of most of the different sounds. It is made up of multiple muscles, and has the ability to perform complex alterations in shape, position and size. Other important articulators are the lips, teeth, alveolar ridge, hard palate, velum and uvula. The lips and the velum can be used in order to alter the shape and size of the oral cavity. The function of the articulator in the production of speech sounds will be discussed in detail later.

## Phonation

Phonation refers to the function of the larynx and the different uses of vocal fold vibrations in speech production. It is common to divide the vocal fold settings into different modes, and combinations of these, in order to describe the way we use the voice. Differentiation between these is also important in speech analysis, as the different modes have different acoustical properties. In normal English speech only two of these modes are applied: voiced and unvoiced sounds. Other modes can be used as an attempt of voice disguise or be the result of a speech disorder.

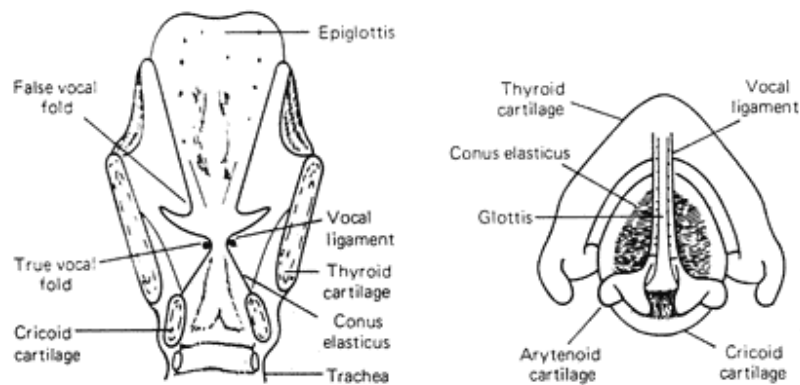


Figure 3: The larynx seen from a) the back b) above.

### **Voiced Sounds**

A normal voiced sound occurs when the arytenoids keep the glottis is closed, in order to create vocal fold vibrations. These vibrations occur along most of the length of the glottis, and their frequency is determined by the tension in the vocal folds. We have the ability to control this tension; something we do in order to alter the pitch during phonetic intonation and singing. Males usually have a frequency range of eighty to three hundred Hertz (cycles pr second), while females can reach five hundred Hertz. Voiced English speech sounds include the consonants [b, d, ɡ, dʒ, m, n, ŋ, v, ð, z, ʒ, l, r, w, j] together with all the vowels and diphthongs.

### **Unvoiced Sounds**

An unvoiced sound is characteristic of its absence of phonation. Movements of the arytenoids separate the vocal folds and the glottis are held open at all times during these types of sounds. The opening, which is between sixty and ninety-five percent, lets the airflow pass through without creating any vibrations, but still accelerates the air by being

narrower than the trachea. The absence of vocal fold vibrations means that unvoiced sounds will not have a fundamental frequency and that we are unable to control their pitch. In English, unvoiced sounds include the plosive and fricative consonants [p, t, k, tʃ, f, θ, s, ʃ, h].

### Alternative modes

There are other phonation modes that should be mentioned, because they, although not component of normal English speech, can be used in artificially attempting to disguise the identity of the speaker: Whisper are created by keeping an opening between the arytenoids while the vocal folds are kept together. This changes the properties of voiced sounds, but leaves the unvoiced sounds unaffected. Creaking, with its characteristically low frequencies, is produced with the glottis closed, but also has settings in the vocal folds that only allow a short fragment of them to vibrate. Falsetto is similar to voiced sounds in that vibrations occur along the whole length of the vocal folds, but gets its characteristically high frequencies by shaping the vocal folds so that they get a thin edge. Murmur is combination of voiced sounds without full closure of the glottis. Its breathy characteristic is caused by the constant airflow allowed through. Combinations of the different modes are also possible.

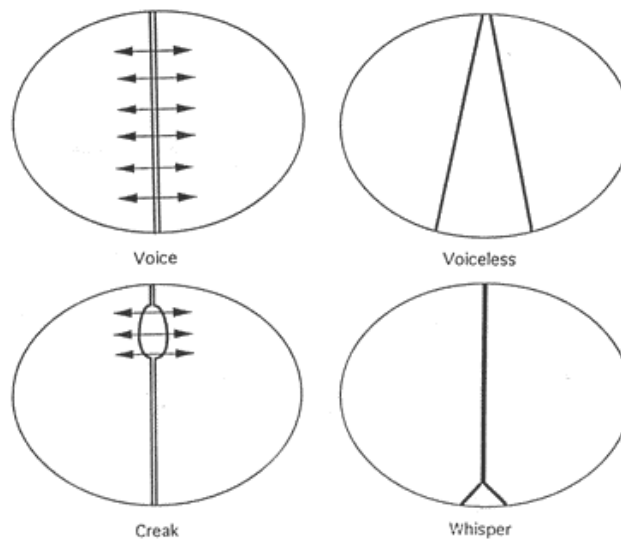


Figure 4: Vocal fold settings in some common modes.

## Articulation

Articulation refers to the ways we use the numerous organs in the upper vocal tract to create the different sounds that form speech, and it includes both the formation and degree of constriction. The two main groups of articulated sounds, consonants and vowels, are defined accordingly to the presence or absence of constriction. The consonants are again divided into subgroups regarding the manner in which the constriction is created. Each of the groups and subgroups represents different auditory events and will accordingly be described separately. There are more sub groups of consonants than the ones mentioned below (e.g. flaps and trills), but as these are not used in normal English speech they are excluded from further description in this paper.

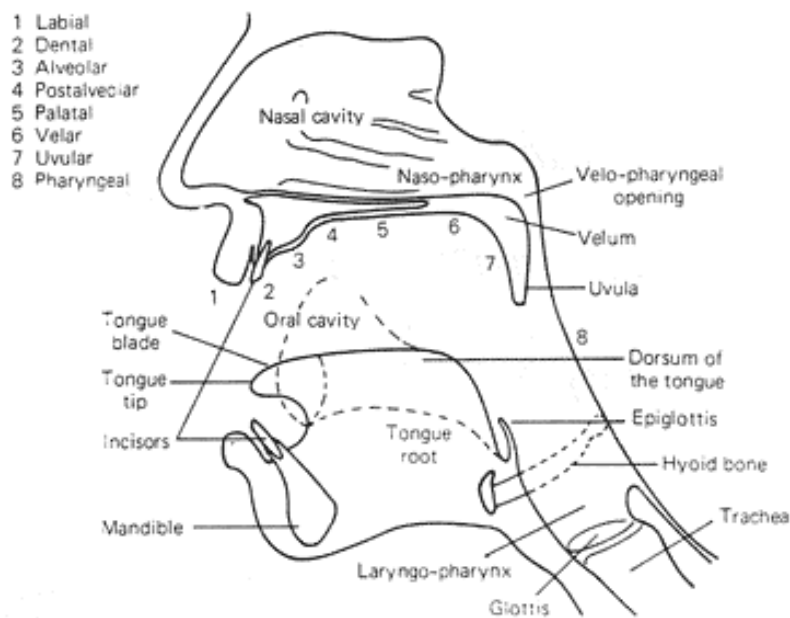


Figure 5: The vocal cavities and sights of articulation.

## **Vowels**

Vowels are voiced sounds formed with an open vocal tract, and their phonetic contents depend on the size and shape of the resonating cavities. All the vowels in English speech are oral (formed in the oral cavity with the velum raised) but nasalized vowels are also possible and are used in other languages such as French. In describing vowels we define them according to the positioning of the tongue and the shape of the lips. In the cardinal system, sixteen reference vowels are defined as the outer limits of sound possible to create without any constriction in the vocal tract. These are divided into two sets. The eight primary ones are related to the position of tongue in a two dimensional space, with axes describing horizontal and vertical placing in the oral cavity. The secondary set has the same tongue positions as the primary set, but differs in the shape of the lips. Although



## **Consonants**

Consonants are sounds created with a constriction in the vocal tract, and their phonetic contents are dependent on the place and manner in which the articulation is created, combined with the phonation mode applied. The possible positions of articulation range from the lips in front, to the glottis at the back of the vocal tract. Although this is a continuous scale, it is common to divide it into regions related to the articulators involved. These are bilabial (lips), labio-dental (lips to teeth), dental (tongue to teeth), alveolar (tongue to top of alveolar ridge), post-alveolar (tongue to back of alveolar ridge), palatal (tongue to hard palate), velar (tongue to velum), uvular (tongue to uvula), pharyngeal (pharynx) and glottal (vocal folds). Combined with the manner of articulation, which involves the level of constriction and movements of the articulators, this leads to a set of sounds that can be both voiced and unvoiced. Not all of these potential consonants are possible to pronounce (e.g. voiced glottal) and only a few are used in English speech.

## **Plosives**

Plosives are sounds created by the formation and release of a full closure in the vocal tract. During the closure, which is heard as a short period of silence, pressure is built up behind the articulators. When the closure is released by separating the articulators, equalization of the pressure differences result in the short sound characteristic of plosives. The duration of both the closure and the release will be dependent on the phonetic contents of the sound and the connection in which it is used, but will otherwise be relatively consistent, as it is impossible to extend them without making a full stop in the word. Plosives in English speech can be both voiced and unvoiced, and the ones used are [p, b, t, d, k, g].

## **Nasals**

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. There are normally no constrictions of the airflow inside the nasal cavity, and since it is impossible to alter its resonating characteristics, the phonetic contents of nasals are entirely determined by the size and shape of the oral cavity behind the closure. The nasal consonants used in English speech [m, n, ŋ] are all voiced, but unvoiced nasals can be produced, although rare because of their low intensity.

## **Fricatives**

Fricative consonants are created by articulation settings that have a constriction in the vocal tract, narrow enough to make turbulence in the airflow. It is common to divide

these sounds into sub groups as some of them have sibilant (hissing) characteristics [s, z, ʃ, ʒ] from the noise components of the turbulence. Movements in the articulators must also be considered in order to differentiate between sounds that are created at the same location. The ones that have movements are called affricates [tʃ, dʒ], and can be described as “plosives released into fricatives”. Fricatives can be both voiced and unvoiced, and the ones used in English speech are [f, v, θ, ð, s, z, ʃ, ʒ, tʃ, dʒ, h].

## Approximants

Approximants are consonants created with a greater level of constriction than vowels, but not enough to create turbulence in the airflow. It is also common to divide these into subgroups related to the level and shape of the constriction. The ones with the least constriction are often referred to as semi-vowels [j, w], because of their similarity to vowels in articulation setting. Central [r] and lateral [l] approximants refer to the shape of the tongue, where the latter has an opening for the airflow on each side. Approximants are normally voiced sounds, and the ones used in English speech are [l, r, j, w].

	Bilabial	Labio-Dental	Dental	Alveolar	Post-Alveolar
Plosives	p b			t d	
Nasals	m			n	
Fricatives		f v	θ ð	s z	ʃ ʒ tʃ dʒ
Approximants				l r	

	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosives		k g			
Nasals		ŋ			
Fricatives					h
Approximants	j		w		

Table 1: Articulation sights of standard English consonants.

## **The Nature of Sounds**

In analysing speech sounds it is the properties of the sound, and not the sound source itself, that is the subject of study. In understanding both the transit from phonation and articulation to sound and the way sounds are analysed, some basic knowledge about its properties are of importance. Acoustics is the scientific name for the behaviour of sounds waves, which closely relates to standard wave theory as it is described and applied in physics. I will in this section give a brief introduction to the concept of complex waveforms, frequency, amplitude, filters and resonance.

### **Frequency and Waveforms**

Sounds are pressure waves transmitted through a medium. Although sound can travel through many different mediums, with speech sounds this is normally air. In its simplest form, sounds can be described as sinusoidal waveforms plotted on a graph with pressure and time as accordingly the vertical and horizontal axis. One cycle of this waveform completes itself when its displacement (pressure) has returned to its original position. Frequency of the wave is defined as the number of cycles completed over a certain period of time. The standard way of expressing frequency is in Hertz (Hz), which has the unit of period's per second. Each sinusoidal waveform represents one single frequency. When several frequencies are represented in one waveform, their sinusoidal components are added together to produce what is called a complex waveform. A periodic complex waveform has a fundamental frequency, which we perceive as the pitch, together with its harmonic components. The harmonic frequencies are always multiples of the fundamental frequency, so a periodic waveform of 100Hz will have its first harmonic frequency at 200Hz, the second at 300Hz and the third at 400Hz etc. When a complex waveform has a random distribution of the frequency components, the result is a rapidly changing waveform without any general pattern. This is called an aperiodic complex waveform, and is what we commonly perceive as noise. At the other extreme in complexity from sinusoidal waveforms, we have what is called white noise. This is a waveform where every possible frequency is randomly present.

### **Amplitude and Intensity**

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. In complex waveforms it is common for the different sinusoidal frequency components to have different amplitudes, and harmonic frequencies are often characteristic of having decreasing amplitude as their frequencies increase. In describing the volume of a sound it is convenient to utilise the term intensity, which refers to the energy of the waveform. This is equivalent to the waveforms average degree of displacement, and can be derived from the amplitudes as the root mean square (RMS) value (70.7 percent of the peak value in sinusoidal waveforms). The advantage of

this approach is that it can express both the intensity of single frequencies as well as the overall loudness of a complex waveform. To discriminate between the two variables it is normal to refer to the amplitude, implicitly meaning its RMS value, when it comes to the intensity of a specific frequency, while referring to loudness as the overall intensity of a complex waveform. The standard way of expressing intensity is in decibels (dB). This is a logarithmic scale developed to compensate for our perception of differences in intensity as differences in volume. It is a relative scale, meaning that it always compares the intensity to a reference point. This point is usually the threshold of hearing when it comes to loudness and the maximum peak value when it comes frequency analysis of amplitude.

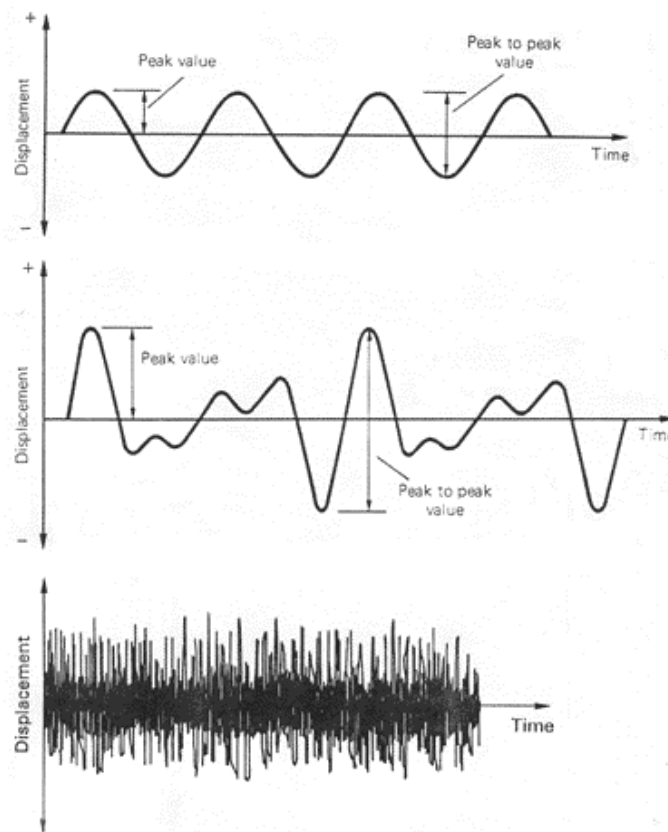


Figure 7: Examples of sinusoidal, periodic complex and aperiodic complex waveforms.

## Filters

A filter is the term used to describe any object or device that can alter the amplitudes of a complex waveform as a function of its frequencies. These can be of any nature dependent on the physical properties of the waves in question, although the only ones that will be discussed in this context are either “acoustical” or “electronic”. There are basically two different kinds of filters along with a variety of subtypes that can be derived from their combinations. A lowpass filter is characteristic of suppressing high frequency

components while letting the low ones through unaltered. A highpass filter is considered the opposite as it suppresses the low frequencies, while letting the high ones through. One combination of these two filter types will result in a bandpass filter, as both the frequency components above and beneath a specific range of frequencies will be suppressed. The frequency components that are let through this type of filter are determined by the upper and lower limits of this range, and the distance between these are normally referred to as the bandwidth. When several bandpass filters are combined, the result will be a multiple-bandpass filter that is characteristic of letting specific bands through, while suppressing the frequency components in-between.

## Resonance

In understanding resonance, it is important to differentiate between free and forced vibrations. Free vibration is when a system is set into vibrations by an initial and final input of energy. These vibrations will always occur at a specific frequency, called the natural or resonating frequency, and their value is determined by the physical properties of the system. Forced vibration occur when there is a constant input of energy into the vibrating system, and if vibrations from another system are the source of this input, the forced system will start to vibrate at the same frequency. Resonance is what occurs when the forcing vibrations have the same frequency as the natural frequency of the resonating (forced) system. The amplitude of the vibrations in the resonating system will in this case have its maximum possible value and thereby cause an amplification of the vibrations. If the forcing vibrations have a higher or lower value than the resonating frequency, this amplification will decrease along with the difference between them. When it comes to sound waves, the nature of resonance becomes more complex as more frequencies are involved. An acoustic resonator normally has multiple different resonating frequencies, with values dependent on its shape and size, and the forcing system is often the pressure waves themselves containing a variety of harmonic frequencies. The resonator will, in this context, possess the same properties of a multiple-bandpass filter, amplifying only the selection of harmonic frequencies that lie inside the bands, centred on the resonating frequencies. When applied to speech sounds these bands of resonated frequencies are called formants and their centres (the resonating frequencies) the formant frequencies.

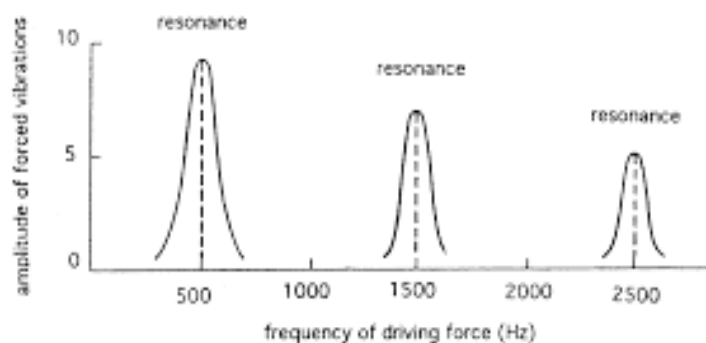


Figure 8: Resonance response of a system with three resonating frequencies.

## Acoustics of Speech Sounds

A simple way of describing the acoustic functions of the vocal organs in speech production is to view them in terms of a source and filter model. There are two different sound sources used in speech; vocal fold vibrations and air flow turbulence. Phonation creates a periodic complex waveform, while turbulence is the source of an aperiodic complex waveform. These are then filtered by the frequency selective resonating cavities before the output is released as speech sounds. This simple model is both an idealisation and a simplification, useful in describing the different acoustical makeup of the different speech sounds. The real world is more complex as none of the sources or filters are ideal.

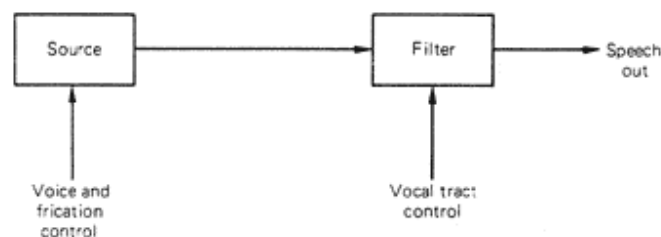


Figure 9: The source and filter model of the acoustics of speech production.

## **Formant Structures**

Since the pressure differences in the airflow are proportional to variations in the glottis size, they will follow the same waveform as the vocal fold vibrations themselves. As one cycle in the vocal fold vibrations will consist of both a period of full blockage together with considerable discrepancies in both release and closure time, this waveform will be of a periodic complex nature. Its harmonic contents are dependent on the duration of these three cyclic phases, and will vary with the mode of phonation and the amount of  $P_{sg}$  applied. When the airflow passes through the filters in the resonating cavities, some of the harmonic frequency components will be amplified, creating peaks in the frequency spectrum. These peaks are the formants and their frequencies will depend on the resonating characteristic of the cavities. These are determined by the manner and position of articulation, so the structure of the formant frequencies will vary with each sound.

A speech sound created solely by vocal fold vibrations as the sound source, will have its phonetic contents mainly determined by the three first formants: F1, F2, F3 and the relative distance between them. The fundamental frequency is, in this context, normally labelled F0. Vowels have in their steady state a static formant structure with phonetic contents determined by variations in F1 and F2, while F3 remains relatively stable. In male speakers F1 and F2 will be in the regions 250Hz to 800Hz and 800Hz to 2.5kHz, while F3 lies at about 2.5kHz. F2 will commonly decrease if we move from the front to

back vowels, while F1 will commonly increase if we move from upper to lower vowels in the cardinal diagram. Females and children will have higher frequency values, but the overall patterns will be the same.

The nasal consonants [n, m, ŋ] are special in the way that they use both the nasal cavity and a closed oral cavity as filters. The formants are mainly caused by resonance in the nasal cavity, so the different sounds will therefore have similar static formant structures. The nasal formants generally have lower intensity than the oral ones, and occur in the regions 250Hz (F1), 1kHz (F2) and 2kHz (F3). The acoustic differences between the nasal consonants are caused by the resonance in the closed oral cavity. This introduces an additional filter, variable in the region of 800Hz to 2kHz dependent on the position of articulation, which selectively decreases the amplitude of the corresponding formants.

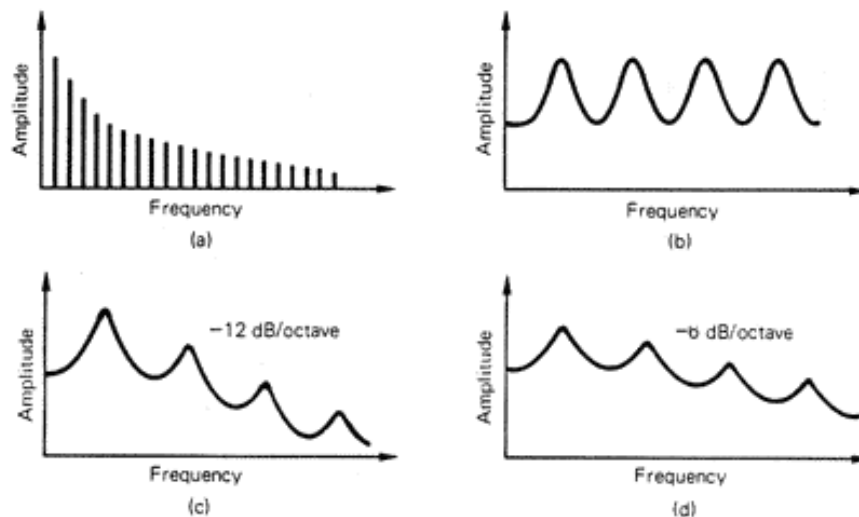


Figure 9: Acoustic properties of formant structures: a) harmonic frequency spectrum of the phonation sound source, b) resonance/filter response of the vocal cavities, c) frequency spectrum of the resonated/filtered sound, d) frequency spectrum of the radiated pressure waveforms.

## Formant Transitions

Movements in the articulators between different sounds will cause smooth frequency transitions between static formant structures. This is an important part of continuous speech as it appears between almost every single sound in a word, but it is also a characteristic part of some types of sounds. Diphthongs have a long first vowel followed by a formant transit to the static structure of the short end vowel. This transit is a glide through all the possible vowels, represented by a straight line on the cardinal diagram, between the two static structures. Diphthongs are therefore characteristic of having continuous movements in both F1 and F2, between the first and the second static vowel structure, while F3 remains stable.

Approximants are also characteristic of having formant structure transitions. These have a short articulation followed by movements towards the next sound. The semivowels [j, w] have a similar starting format structure as the upper vowel, with only a difference in F2 relating to the position of articulation. The central approximant [r] also has this similarity with the upper vowels in the two first formants, but is characteristic of having a very low F3. The lateral approximant [l] has low values in both F1 and F2. Central and lateral approximants normally have a faster transition to the next sound than semivowels.

## Noised Sounds

Turbulence occurs when there is a constriction in the vocal tract narrow enough to cause friction for the airflow. Normally, this is related to the manner of articulation, but it can also be a mode dependent part of phonation. Articulated turbulence, as a sound source, presents an aperiodic complex airflow, which is dependant upon the cross section of the constriction. This airflow is then filtered by the cavity between the lips and the place of articulation that, being dependent on its length, will only transmit a certain band of the frequencies. The overall speech output will in this case be a noised sound with frequency components in the whole region of the frequency band.

The range of the frequency band in fricatives will be characteristic of the position of articulation. This will, together with the presence or absence of a fundamental frequency, determine their phonetic contents. The frequency bands in the fricatives used in English speech are approximately 6kHz to 8kHz for labio-dental [f, v] and dental [θ, ð], 4kHz to 8kHz for alveolar [s, z] and 2kHz to 6kHz for post-alveolar [ʃ, ʒ, tʃ, dʒ]. Voiced fricatives normally have less intensity and a narrower band of noise components than their unvoiced partners. These should theoretically also have a formant structure, but due to the intensity of the noise components, this is normally not detectable.

The phonetic contents of plosives are determined by the short burst of noise and the following formant transitions to the next sound. The burst normally starts with a very short peak over a large frequency range before the position of articulation restricts the band. The frequency bands after the peak are, in the English plosives, distributed with ranges of 600Hz to 800Hz for bilabials [p, b], around 4kHz for alveolar [t, d] and from 1800Hz to 2kHz for velar [k, g]. Another characteristic acoustic property of plosives is the duration of the different phases of articulation. The silent part can last from 70ms to 140ms, while the bust can reach from 10ms to 50ms. The lengths of these phases are dependent on the position of articulation, and the voiced plosives are usually shorter than the unvoiced.

## *Part 2:*

### **The Theory of Voice Identification**

As described in the first part of this paper, the nature of speech sound is dependent on three properties: the dimensions of the vocal organs, the mode of phonation and the manner of articulation. It is the individual differences in these that make discrepancies between different peoples voices and allows us to recognise audible, a particular person in conversation and visually, through forensic voice identification. The size and shape of the vocal cavities are properties that remain relatively consistent, although they can undergo minor alterations dependent on ageing and physical health. Their measurements are highly characteristic of an individual. Some of the contributions these make to speech sounds will therefore be good markers that may be used for discrimination when applying forensic voice identification. Phonation and articulation are something we learn by listening to other speakers as infants. Through trial and error we adopt a habitual pattern of both, settings in the vocal folds and movements in the articulators, which will be reflections of our environmental influences during this learning process. Although the overall pattern will be similar in every person speaking the same language, this being what determines the phonetic contents of an utterance, there exist small individual differences that cause one speaker to sound different from another.

### **Interspeaker Variations**

Interspeaker variation is the term used to describe all differences in the acoustic parameters of speech sounds (caused by vocal organ dimensions, habitual phonation and articulation patterns) that can be used to differentiate between speakers. This is a concept that has not been completely explored yet, as we still don't know which acoustic parameters the ear and the human brain use when distinguishing between speakers. Some parameters, however, are well defined and measurable through analysis. The generalised formant structures and energy distribution of the different phonemes, as described in the first part, are produced by all speakers and can be used to determine the phonetic contents of speech. In frequency-based analysis of sound, this will cause a similar overall pattern in two utterances of the same word or phrase by two different speakers, but individual differences will be present and can be detected. Mimicry is an attempt to minimise interspeaker variations so that listeners will perceive two separate utterances as coming from a single person when they actually come from two. Although this can be successful when heard, it is normally possible to differentiate between the speakers through analysis, as it is very difficult to alter every parameter involved. Which, of the available acoustic parameters, one chooses to concentrate on during voice identification is dependant on the method of comparison. Since the three different available techniques use three different sets of parameters, these will be dealt with separately, later in this paper.

## **Intraspeaker Variations**

A major problem encountered in voice identification is the fact that people are not consistent in the way they speak. Although every person has a habitual user pattern of phonation and articulation, there will be small discrepancies in each utterance of the same word or phrase during normal speech. The effect of this is that when one person utters the same word twice, they will not be acoustically identical although a listener perceives them as the same word. Intraspeaker variation is the term used to describe the acoustic events that are the results of this inconsistency in the user patterns of the vocal folds and the articulators. Since this unfortunately affects many of the same acoustic parameters that are used to detect interspeaker variations, it is fundamental in accurate voice identification, to have the ability to distinguish between these two types of variation. There are a lot of different events that are known to increase intraspeaker variations. The most important ones that vary during normal speech are time and the context in which the speech take place. People undergo small changes in speech patterns over time, making identifications based upon samples taken at very different points in time difficult. The exact utterance of a word will be influenced by the preceding and following words, together with the overall intonation of the sentence, causing identifications based on off contextual word samples to be difficult. Other explanations for intraspeaker variations, are changes caused by emotional state, deliberate disguise and the influence of drugs or alcohol.

## **Scientific Basis**

The scientific basis of voice identification relies on the assumption that intraspeaker variations are less prominent or different to interspeaker variations. This presents us with a fundamental problem in the scientific validity of forensic voice identifications, as there have been a very limited number of studies conducted in order to understand the origin and the characteristics of the different kinds of variation. The relation between inter and intra speaker variations is still not well defined and most of the studies done on voice identification have used parameters that have been judged, by the experimenter, likely to be most characteristic of a person. This lack of scientific basis for a theory can be replaced by experimental studies and statistical evaluation of the different parameters. This, however, requires an extensive amount of experiments with uniform conclusions based upon a large database. As will be discussed later, this is not the case with forensic voice identification as it is performed at the present moment.

## **The Sound Spectrograph**

The spectrograph is an instrument used to analyse the complex waveforms of sound and their alterations in time. This is done through spectrograms, which are graphic displays of the amplitude as a function of both frequency and time. The spectrograph was first developed by the Bell telephone laboratories in a governmentally founded study during World War II, and presented by Koenig, Dunn and Lacy in 1946<sup>1</sup>. It has since, gone through many technical improvements, but the basic principle is still the same. The spectrograph is today, standard equipment in most speech labs, where it is used in analysis and classification of speech sounds and in treatment of speech and hearing disorders. In forensic science it has found its use as the primary tool used to identify people by their voices.

### **Analogue Systems**

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 is 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. The finished result, the spectrogram, is a two dimensional plot with time and frequency as the horizontal and vertical axes. The differences in amplitude values are shown in a grey scaling where black represents the most intense and white the least intense waveform components.

### **Accuracy and Resolution**

The length of the speech sample able to be displayed on a spectrogram is dependent on the size of the drum. The standard drum size corresponds to a 2.4s long sample, although larger drums are available. The operator can adjust the frequency range of the total spectrogram according to the parts of the frequency spectrum that are of interest. It is common to use a frequency range of 0Hz to 8kHz, as this range contains most of the phonetic contents of speech. However, other circumstances such as the quality of the recording can influence this decision. The sensitivity of the paper sets limits as to the differences in amplitude able to be displayed on the spectrogram. Since this range, approximately 12dB, is much smaller than the variations of interest in speech sounds, the signal is usually electronically compressed before printing. This means that a larger amplitude range is linearly decreased to fit inside the sensitivity range of the paper. Although this can be used to display large variation of intensity, it will at the same time

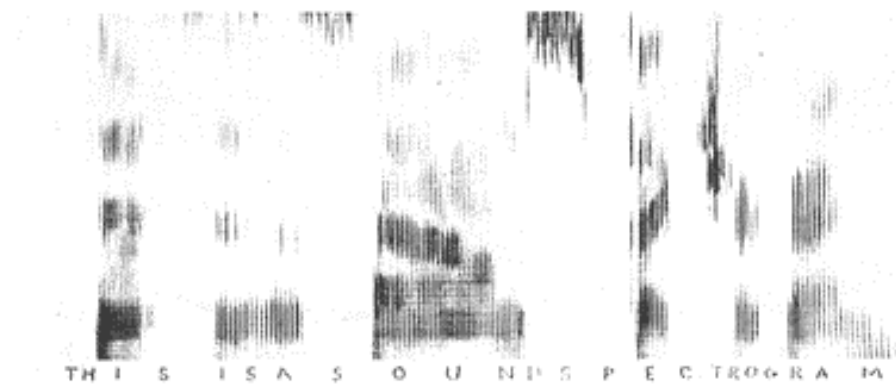
decrease the amplitude resolution of the plot. In order to limit the amount of compression required to visualise the less intense upper formants, it is common to use a frequency dependent amplification of the signal before plotting. This is normally done with a rate of increase at 12dB per octave (doubling of frequency).

## **Filter Modes**

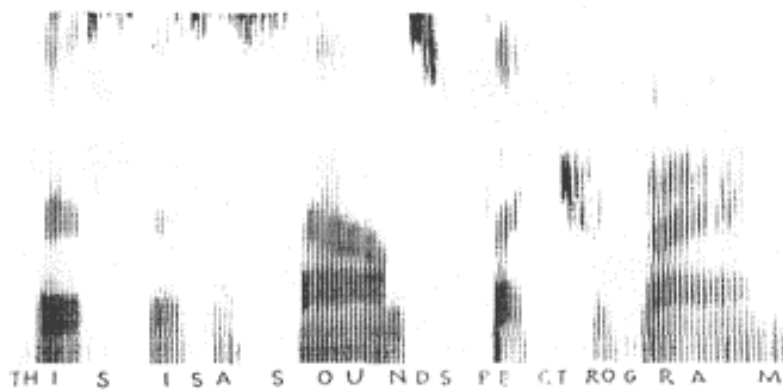
The stylus will print the overall energy of the frequency band that is passed through the filter, making the frequency resolution in the spectrogram dependent on the bandwidth of the filter. The different frequencies inside this band cannot be visually separated in the spectrogram, so the narrower the bandwidth of the filter is, the more detailed the frequency information will be. Another important property of the filter is that it has a delay in the response time, which will influence the time resolution of the plot. If this delay increases, there will be a corresponding decrease in the time resolution. The result of this is that information about rapid alterations in frequency will be lost. Since the response time of a filter is dependent on its bandwidth (approximately the inverted value) it means that there is a direct trade-off between time and frequency resolution in a spectrogram. A narrowband filter (45Hz bandwidth with 20ms responsetime) will produce a plot that is useful for analysis of the harmonic frequencies inside the formant structures, as these are visible as separate bands. However, it will be difficult to use in analysis of speech patterns, as information about short duration sounds, will be lost. Since the overall speech patterns are usually of more interest than exact frequency information, a broadband filter (300Hz bandwidth with 3ms responsetime) is normally utilized. This filter will smear the individual harmonic frequencies inside the formants, but the overall formant patterns will be visible as large bands in the plot, together with all the short duration sound that is important for the phonetic contents.

## **Digital Systems**

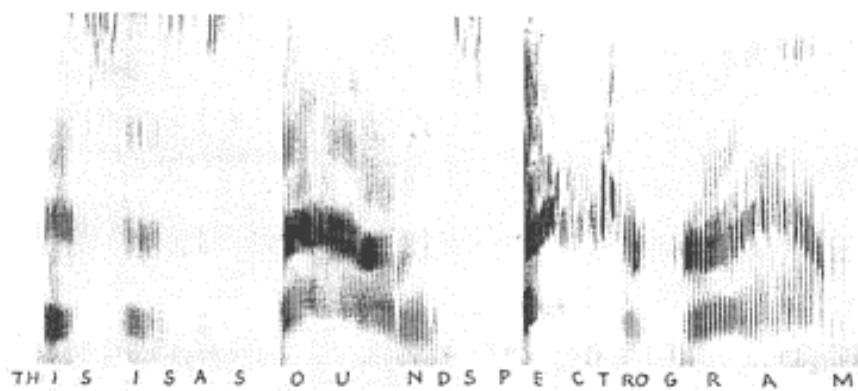
The invention of computerized sound processing has lead to the development of digital spectrographs. These can in design range from digital sound processing workstations to PC based software systems. The digital spectrograph accomplishes the same tasks as the analogue, but offers a lot more flexibility in the developments of the spectrograms. It is possible to display multiple plots at the same time and alter their frequency range, time alignment, amplitude sensitivity and filter bandwidth simultaneously. These spectrographs also use high quality sound and do not place any restrictions on the length of the speech samples. The computerized speech analysis systems, including the spectrograph, normally also contains a set of mathematical tools, which can assist the examiner in making more numerical analysis of speech. As these numerical analyses can be more discriminating than the actual spectrogram themselves, although not used as a standard in forensic voice identification, some of them will be discussed in the last part of this paper.



*Speaker A: Day 1*



*Speaker A: Day 4*



*Speaker B: Day 1*

Figure 11: Three spectrograms of the phrase “this is a sound spectrogram”. The two first displays intraspeaker variations and is an example of the basis on which voice identifications has to be performed. The third spectrogram displays interspeaker variations and is an example of the basis on which voice elimination has to be performed.

## **Spectrographic Identification Studies**

A spectrogram is an indirect graphic representation of the habitual user pattern of phonation and articulation. The formant structures and their transitions will be visible as a set of three or more horizontal dark lines with relative vertical movements in time, while noised sounds will be visible as dark fields inside the corresponding frequency band. If one applies the information on the general acoustic makeup of speech sounds, as presented in the first part of the paper, it is possible to read the phonetic contents directly from the spectrogram. Interspeaker and intraspeaker variations in speech patterns can be acoustically analysed through spectrograms, a discovery that led to research on using these in order to establish the identity of the speaker.

### **Kersta's Voiceprinting Method**

Kersta first developed the idea of identifying people on the basis of their spectrographic plots in 1962<sup>2</sup>. On a request from the New York City Police Department, he developed a method called voiceprinting. This involved comparison of ten cue words, commonly used in English speech, delivered from the unknown speech sample together with known reference samples. His claim as to for the scientific validity of this method, was that the possibility of two persons having identical dimensions of the vocal cavities and user patterns of the articulators seemed very remote. To support this claim he conducted experiments to see how well test examiners could identify persons from matching of their spectrograms. Eight female high school students were given one week of training before they were to conduct closed set experiments of identification. Closed set refers to experiments in which the unknown sample is to be identified out of a set of reference speakers, and the examiner is told in advance that a match exists. This is opposed to open set experiments where a match does not necessarily exist. In his experiments, the number of reference speakers varied in size from nine to fifteen and the examiners were told to identify all of the cue words separately. The false identification rate for the different words had an average of 1%, and it was supposed that if several words were used together in the identification, the overall error rate would be the product of the error rates of the individual words.

Kersta's voiceprint method received a lot of attention from both media and the scientific community, due to the low reported error rates and its many similarities to the successfully applied method of fingerprint identification. Its applicability in forensic science was promising as it offered a method of seemingly objective identification together with the possibility of a classification and filing system. Such a system was soon to be developed by Kersta and his co-workers. The response from the scientific community was divided. Some of the objection was due to the fact that the construction of the experiments was not relevant to real forensic situations and that voiceprinting could not be compared to fingerprinting. Visible intraspeaker variations in the spectrograms would cause plots of two utterances of the same words to not be as identical as fingerprints<sup>3</sup>. In his published results Kersta only used isolated utterances of the cue

words for both the unknown and the reference samples, stating that differences in the results from experiments where the words had been taken from contexts were minimal. To check this statement Young and Campbell conducted a small experiment in 1967<sup>4</sup>. They reported error rates of 21.6% when comparisons were made between words spoken in isolations and 62.7% when the words were taken from a random context. Stevens *et al* conducted another small study in 1968, in order to compare the reliability of spectrographic identification to aural identification<sup>5</sup>. They reported an error rate of 21% for spectrograms, while aural identification was found more reliable with an error rate of only 6%. These studies were subjects to much of the same critique as Kersta's, and the conflicting results were thought to arise from different experimental procedures. This view was based upon the fact that a review of the work performed by some of Kersta's trainees<sup>6</sup> supported the 1962 results, as an average error rate of 5.7% was reported.

### **Michigan State University Voice Identification Project**

The first and only major study on the subject of spectrographic voice identification, was conducted by Dr. Tosi and a group of speech scientists at the Michigan State University from 1968 to 1970<sup>7</sup>. This was done in conjunction with the Michigan Department of State Police and sponsored by the U.S. Department of Justice. The purpose of the study was to investigate a wide range of variables concerning spectrographic identification and to see how these corresponded to previous studies and real forensic situations. For the voice samples, 250 persons with no distinct dialect differences or speech disorders, were selected randomly from a population of 25 000 male university students. These speakers then went through two different recording sessions where speech samples were collected under a variety of conditions, including both recording equipment and the context in which the cue word was spoken. Twenty-nine people were hired as test examiners and given one month of training. The different test settings in the experiments included the use of different amounts of cue words, number of speakers in the reference population and reference prints of the same cue words. These settings were then used to investigate how different recording conditions, contextual and time variations of speech would influence the result of the comparisons. Both closed and open set experiments were performed where the examiners were forced to make a decision while at the same time indicating their degree of confidence. Every possible combination of these variables and their different levels were tested with a total of 35000 attempts of identifications performed. Some of the most interesting results from this study included the closed set experiments where nine cue words spoken in isolation at the same recording sessions were used as both unknown and reference samples for the comparisons. This was basically an identical experiment to Kersta's and it supported his findings, as the average false identification rate was 0.5%. The forensic models presented in this study were open set experiments, including nine cue words recorded on different occasions and spoken in both fixed and random contexts. In these experiments there are two possible errors: false identification (the examiner selects a wrong match when either a correct match exists or not) and false elimination (the examiner fails to recognize a match that exists). When the cue words for both the unknown and the reference samples were taken from identical sentences, the false identification rate was 4.2% and the false elimination rate 10.1%.

When the cue words were collected from different contents these error rates were 6.4% and 11.8% respectively. From the confidence rating of the decisions, it was calculated that if the examiners were not forced to make a decision, a positive conclusion would only be made in 74% of the identification attempts, with approximate error rates of 2% for false identifications and 5% for false eliminations.

Some of the criticism of the Michigan State University Voice Identification Project regarded the validity of the results obtained by the forensic models<sup>8,9</sup>. In a real forensic situation the examiner will be in the position where he/she is to compare the unknown sample to reference samples from only one or a few suspects. In these situations the size of the reference population will be infinite. Since the MSU study showed significant increase in error rates as the reference population was enlarged, it was suggested that the reported error rates should be viewed upon as artificial minimums. This claim was also supported by the fact that the study showed an increasing error rate when intraspeaker variations were considered. Factors likely to be encountered in real forensic situation such as emotional state and attempts of disguise was not covered by the study, and these were considered as factors that would further decrease the accuracy.

## **Voice Disguises**

To address some of the problems regarding intraspeaker variations, Reich conducted a study in 1975 on the effect of selected voice disguises<sup>10</sup>. This study followed a similar scheme as the forensic models in the MSU study, but was considerably smaller as only four test examiners and a population of thirty speakers was used. The four test examiners were Ph.D. students in speech pathology, who received one month of training. In this study the examiners were to compare samples of the cue words produced in both undisguised and disguised utterances to undisguised reference samples, in an attempt to see how this would affect the accuracy of the decisions. The different error rates in this study were not explicitly published, but general performance was not consistent with the results in the MSU study as the overall error rate in the comparisons of the undisguised words was 43.3%. An interesting result of this study is the fact that the general performance was considerably worse when identifications were made from the disguised samples. An average overall error rate of 63.7% in these experiments indicates that disguise indeed increases the intraspeaker variations seen in voiceprints.

## **The Classification and Filing System**

Hazen performed in 1972, another interesting study, in order to test Kersta's classification and filing system of reference print<sup>11</sup>. This system was supposed to account for intraspeaker variability, as the two most different print of the same word were filed together and used in the comparisons. It was also claimed that it would serve as a population reduction tool, as the speaker classification system could be used to obtain a small number of possible suspects from the speakers with the most similar patterns. Seven panels, each consistent of two law enforcement trainees, were trained in Kersta's

voiceprinting method, and were to perform experiments on a population of fifty reference speakers. The experiments was divided into two parts as the examiners was first to reduce the population size to a limited number of suspects, and then do a positive identification or elimination based on this result. When the cue words were taken from a random context (as would be the case in the filing system) an error rate of 52.4% was reported in the task of including the match as one of the suspects. When the examiners were to make positive identifications or eliminations, further error rates of 16.7% false identification and 66.7% false elimination were reported. When one compares this data with the results obtained for the comparisons made with the cue word spoken in the same context (42.9% reduction error, 7.1% false identifications and 35.7% false eliminations) it shows that the filing and classification system is not supported by the study.

## **Forensic Identifications**

When it comes to the application of voice identification, as it is used in real forensic situations, the validity of all of the above mentioned studies have been questioned many times<sup>3, 8, 12</sup>. One of the main objections has been that none of them are consistent with the standards of voice identification that have been used by most law enforcement officers. These standards have been subjects of both review and improvements, but have always been different from the laboratory experiments in three major ways: the method of identification, the different conclusions possible to make and the experience of the examiners. Forensic examiners use a combination of aural and spectrographic identification, opposed to the studies where the decisions were purely made on the basis of the spectrograms. As Stevens *et al*<sup>5</sup> pointed out, the usage of aural comparison of speech samples can in some cases be even more discriminating than the spectrograms. When the two methods are used together the examiner will have a broader basis for his/her opinions, and it is therefore reasonable to conclude that the accuracy of the conclusions will increase. The possible conclusions are also different in real forensic situations, as the examiners must take into account their degree of certainty. In all of the mentioned studies the test examiners were forced to make a positive identification or elimination based on the available information. This is a different situation to the one that forensic examiners are confronted with, as they are only allowed to make a positive conclusion in cases where enough information is available corresponding to a high degree of certainty. As noted in the MSU study, the introduction of examiner certainty is likely to decrease the error rates, while significantly increasing the number of non-conclusive decisions. The training and the experience of the test examiners used in the above studies can be viewed upon as limited compared to the requirements necessary for forensic examiners. The ways in which this could have been influential on the results of the studies, was demonstrated in a study performed by Smrkovski in 1975<sup>13</sup>. He compared the performance of experienced forensic examiner at the Michigan State Police Voice Identification Unit with those of persons with a moderate level of training. The results he obtained were that in some cases where the inexperienced examiners had error ratings of 30%, the professionals performed error free.

## **National Academy of Science**

In order to evaluate the uses of spectrographic voiceprint identification and its reliability as a forensic tool, the National Academy of Science formed a group consisting of some of the leading scientists on the subject<sup>12</sup>. In their report released in 1979, they concluded that at the present time there was not enough information available for judicial or legislative bodies to make judgments on an adequate basis, regarding both the reliability and admissibility of spectrographic identifications. They stated that though some information about the identity of the speaker was available through spectrograms, the assumption that intraspeaker variations are different to interspeaker variations had not been adequately explored. Further long-term studies were recommended and three problem areas that needed solutions were cited. These were accurate measurements in real forensic situations, acceptable low error rates for variables in the particular cases under consideration and a determination made as to whether the nature of the possible error source could be adequately explained to a lay jury. The existing studies were viewed upon as different professional judgments from fragmentary data, rather than objective results obtained in relevant forensic situations.

## **FBI's Performance Survey**

The most recent publication on the issue of spectrographic voice identification, and the only one since the NAS report, is a survey done Koenig in 1986 on the performance of the FBI's voice identification unit in real forensic situations<sup>14</sup>. The FBI uses an internal standard that includes both aural and spectrographic identification. In order to make a high confidence conclusion of identification or elimination, at least twenty words have to be judged by the examiner as very similar or very dissimilar. If this cannot be obtained, the result is labelled low confidence or no decision. When a high confidence conclusion is obtained, this has to be confirmed by one or two other examiners through independent evaluations. In the survey, the results of 2000 comparisons performed over a period of fifteen years were summarised. High confidence conclusions were made in 34.8% of the cases and out of these the reported error rate was 0.3% false identification and 0.5% false elimination. Objections have been raised as to the scientific validity of this survey because of the method used to validate the decisions<sup>15, 16</sup>. When a high confidence identification or elimination was determined, the field investigators were contacted to see if the conclusions were consistent with other evidence. If a confirmation of the identification or elimination was not possible, the case was labelled in the survey as a no decision. Because of the possible errors in this kind of feedback confirmation, the presented error rates must be viewed upon only as possible minimum values.

## **Summary**

From the above review of the studies done on spectrographic voice identification, it seems that the accuracy is just as random as a flip of a coin. The reported error rates

range from 0% to 83.3%. There are some considerations one must take before making such a conclusion, those being the nature of the spectrograms and the design of the experiments. Spectrographic identification is different from fingerprint identification. This is because comparisons of spectrograms involve finding sufficient similarities, while fingerprints always have exactly the same pattern. The individual differences in the spectrograms, caused by intraspeaker variation, make it the examiners task to subjectively determine if a match exists. In taking this stand, the reported error rates do not represent the accuracy of the technique, but the performance of the examiner involved. A conclusion one can then make from the above studies is that training and experience of the examiner are likely to improve the accuracy substantially. However, direct comparisons of the above studies must be done carefully, as there may be major differences in their design. Little is reported about the details in the training of the examiners, how the spectrograms were read or on the parameters involved in the comparisons. With one exception, the presented results are based upon a small database and a limited number of examiners performing a small number of comparisons. The result of random effects in these small studies, have been demonstrated through huge variations in the performance of the different test examiners and different error rates for each speaker<sup>4, 5, 10, 11</sup>. Because there are only a limited amount of studies performed on spectrographic voice identification, the ones that have been performed showing conflicting results, it is difficult to make any conclusion about accuracy. It does however seem from a scientific viewpoint, that the 2% of false identifications and 5% of false eliminations with high degrees of certainty in the MSU study are the most reasonable estimations. The FBI survey demonstrates how standardised procedures in real forensic situations can improve these estimations, when highly qualified examiners use the best equipment available.

## **Voice Comparison Standards**

Standardisation of forensic voice identification procedures and certification of examiners was commenced in 1971, when the Voice Identification and Acoustic Analysis Subcommittee of the International Association for Identification (IAI) was formed. IAI stopped their certification program in 1995, but since their standards of examination procedures are still valid, it will serve as the example in this paper<sup>17</sup>. Other associations and organisations have partially or completely adopted this standard<sup>18</sup>. The FBI's internal standard<sup>19,20</sup> is very similar to IAI's and since this is the one used in one of the most important reliability references<sup>14</sup>, comparisons between the two will be made regarding to where they are different.

### **Examiner qualification**

The IAI's requirements for formal education of voice examiners are a High School Diploma with at least one additional course in speech science. A university degree with papers in related subjects such as electronic engineering, physics and linguistics are additionally recommended (required by the FBI). Every examiner must attend a short course in spectrographic interpretation before an apprenticeship period of usually two years commences. During this period the trainee is under the supervision of an examiner that has been certified for at least two years, with a minimum of one hundred comparisons in real forensic cases, resulting in at least twenty-five high confidence level conclusions. When this apprenticeship period is completed, the examiner goes through practical, theoretical and hearing tests every three years in order to be formally certified.

### **Reference Samples**

The quality of the reference samples is critical in making an accurate analysis. Although these samples can be collected by the investigative officer, it is recommended that this is done by the examiner whenever possible. The quality of the unknown sample is dependent upon the recording conditions, which include such factors as the quality of the microphone, transition line and recording equipment. As these factors can to some extent alter the available frequency information, it is important that some of them are emulated in the reference samples. As an example: if the unknown voice sample is recorded over the telephone, this must also be replicated with the reference samples with approximately the same distance between sender and receiver. The only things that must not be emulated in the reference samples are acoustic background noise and the quality of the recording equipment. For this purposes only quiet rooms and high quality tape recorders or digital sound systems are allowed. The speech samples, should ideally, be spoken in the same manner as the unknown voice. In order to achieve this it is recommended to have someone to recite the phrases in the same manner as the unknown sample, and have the suspect to repeat them in a similar fashion. If the suspect is cooperative it may be sufficient to have him/her to read in a normal voice from a familiar transcript. Unless the

whole text is very long, every phrase in the entire text must be repeated until three satisfactory reference samples are obtained.

## **Exclusion of Samples**

The examiner must perform a preliminary examination of the samples in order to see if they satisfy the demands of quality. With few exceptions, only the original recording of the samples can be accepted for the analysis. If these have any kind of distortion or lack of frequency information beneath 2kHz (2.5kHz in the FBI standard), they must be excluded from further examination. However, enhancement procedures can be used in order to remove noise and improve the frequency range of the samples. The examiner must, through aural inspection, determine if the samples contain deliberate attempts to disguise or reveal influences from psychological state or any control substances. If this is the case one should be careful in performing any analysis at all, as this is likely to increase intraspeaker variations to the extent where no final conclusions can be made. An exception to this is if an attempt to disguise is successfully replicated in the reference samples. If the samples meet these requirements, the examiner must compare each word and phrase for intraspeaker variations. Only words pronounced in a similar manner in both the unknown and reference samples should continue to be used in further analysis.

## **Aural Comparisons**

Aural comparisons between the samples are performed as short-term memory examinations. This is done by having different playback units for each sample, with the ability to rapidly switch between them. The first things one will look for during aural comparison of the samples are huge inconsistencies such as sex and age that will exclude further examinations and result in a positive elimination of the suspect. In forensic examinations this is not likely to be a real issue, as the investigative officer on the case will perform this initial screening, and only the difficult cases will be submitted. When extreme dissimilarity does not exist, the examiner must carefully examine the samples and the individual words for components of speech that could be characteristic of the individual. The usual things to listen for during aural examinations are information about: pitch, intonation, stress/emphasis, rate of speech, mode of phonation, vocal quality and possible speech disorders. Pitch and intonation refer to the tone of the voice and the way in which this is altered through continuous speech. This can be useful in distinguishing between speakers, but the examiner must be aware that it is likely to be different if the samples are obtained through conversation or reading. These circumstances are also likely to influence other characteristic parameters, such as the rate of speech and the stress within a word or sentence. Obtaining information about these either from the officer who collected the samples, or purely through listening, are therefore essential. The presence of speech disorders and alternative modes of phonation will add important information to the identity of the speaker, as these can be highly characteristic of speech abnormalities. As mentioned earlier, the ear and the human brain have the excellent capability, not yet fully understood by science, to recognise different voices. The overall

subjective judgement made by the examiner on the general similarity or dissimilarity between the samples will therefore also be of importance to the final conclusion.

## **Preparation of Spectrograms**

A spectrogram of every word used in the aural comparison must be developed. These must, without exceptions, be prepared by using a broadband filter with a bandwidth between 250Hz and 300Hz. The IAI standard states that the frequency range of the spectrogram must be chosen so that the whole frequency range of the recorded speech sample is graphically presented. Here, the FBI uses a standardized range of 80Hz to 4kHz, which corresponds to the bandwidth of telecommunication lines. It is very important that both the unknown samples and all of the reference samples are prepared with the same settings of filter bandwidth and frequency range, as differences in these will result in different plots that will be impossible to use in comparisons. Each spectrogram must, with great care, be marked orthographically, phonetically or both, below each different sound.

## **Spectrographic Comparisons**

A comparison between the reference sample spectrograms must first be performed in order to determine the range of intraspeaker variations for all the parameters. If there is a considerable amount of variation present in a specific word, this must be excluded from the analysis and the final conclusion. When only a limited amount of variation is present in the reference samples, the unknown sample must be examined to see if it is consistent with this range. The approximate frequency, bandwidth and relative intensity of the different formants, together with the shaping and positioning of the transitions must be examined for each word. One of the weaknesses with spectrograms is, as result of the trade off between frequency and time resolution, is the fact that precise values of these parameters cannot be obtained. However, the purpose of the spectrogram is to analyse the shaping, the rate of change and the relative positioning of formants, something it displays perfectly. The energy distribution of the noised sounds must be compared to see if these are consistent between the samples. In samples of poor quality with only limited frequency ranges, this can be difficult as most noised sounds have important components above 4kHz. Similarities in fundamental frequency can be examined through the vertical striations present in the voiced sounds, as these will have timing correspondent to the vibrations of the vocal folds. This can be a very useful characteristic of a voice, but must be used with care when examining separate words, as it is normally altered continuously during intonation of continuous speech. The length of a word can also be characteristic of a voice and can be read off a spectrogram by measuring its length with a ruler. When a person speaks more slowly or faster than normal, it is usually the time between the words that is affected. The effects of inappropriate couplings between sounds can be seen in a spectrogram as either enhanced or diminished energy distribution in a specific frequency band. This is the result of irregular use of the resonating cavities, and results in highly characteristic properties such as nasalized speech. A good pattern match exists, when

most, if not all of these parameters show strong similarities. It is important though to keep in mind that very different voices may exhibit similarities.

## Conclusions

The IAI standards allow the examiner to produce one out of seven possible conclusions for each examination. These represent different confidence levels, which are determined by specific criteria. However, the examiner is allowed to choose a lower degree of confidence for a conclusion even if the criteria of a higher confidence level are met. An *identification or elimination* occurs when at least 90% of the comparable words are judged by the examiner as being very similar or dissimilar through both aural and spectrographic comparisons. The total number of words must be at least twenty, and three or more formants must be visible. These decisions are not allowed if any of the samples contain vocal disguise or are recorded more than six years apart. A *probable identification or elimination* is allowed when at least 80% of the words are very similar or dissimilar both aurally and spectrographically. These must have a total number of at least fifteen words with two or more identifiable formants. If, in this case, only ten or more words can be produced the conclusion will be a *possible identification or elimination*. If the result of the comparisons falls below the criteria of the lowest confidence level the decision must be *inconclusive*. This must also be the conclusion if the samples reveal aural similarities and spectrographic dissimilarities, or vice versa. The FBI standard has the same criteria for the highest certainty level decisions, but do not allow any lower level decisions to be made unless there are at least twenty comparable words. In both standards the high confidence level decisions must be confirmed through the independent analysis of a second examiner.

## Testimony

This is the area in which there is the largest inconsistency between the two standards. The FBI does not allow their examiners to perform any testimony in the court of law and spectrographic voice identifications are only performed as an investigational aid. The IAI standard takes no position regarding this point and leaves the decision up to the individual examiners and the corresponding law enforcement agencies involved. If a decision is made to give testimony, the examiner is obligated to inform the court that the IAI does not consider spectrographic voice identification as being completely conclusive. This kind of accuracy statement must also be added to every written case report.

## **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*<sup>21</sup>. 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”<sup>22</sup>. 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<sup>23</sup>. 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<sup>24</sup>. 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<sup>25</sup>, while others have admitted the evidence even when presented with substantial scientific disagreements<sup>26</sup>. 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<sup>27</sup>. 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<sup>28</sup>. 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.

### *Part 3:*

## **Developments in Voice Identification**

One of the recommendations made by the National Academy of Sciences <sup>12</sup> for possible improvements of forensic voice identification, was the usage of techniques developed for automatic speaker recognition. One of the major benefits of applying these techniques is that the area is still under major developments. While the research on spectrographic voice identification slowly faded out in the late seventies, automatic speaker recognition has been the subject of increasing research ever since the early sixties. All of the different systems designed for automatic speaker recognition are based on open digitally coded waveforms. These codes make up a numerical basis from which detailed information about different parameters of speech can be derived through mathematical algorithms. Since the automatic speaker recognition task is of a somewhat different nature than voice identification, most of the developed systems cannot directly be transferred to forensic situations. It is therefore not in the scope of this essay to evaluate the performance of the different systems developed, but rather to look closer at some of the techniques used, and discuss how the information these reveal can be applied in forensic situations. I will in this part of the paper concentrate on how digital codes of sound can be used to accurately analyse properties in the time and frequency domains of single phonemes.

### **Digital Sound**

Digital sound, as opposed to analogue, refers to the way in which the waveforms are coded inside the electronic equipment used for processing and transmission of the sound information. While analogue is a direct transformation of the waveforms from pressure differences in air to corresponding voltage differences in a conductor, digital systems represent the waveforms through a set of samples with numerical values. The two important properties that describe the digital coding are sampling rate and the number of bits. The sampling rate is a measurement of the timing between each time the waveform amplitude is represented with new samples. It is normal to describe this property in the unit of frequency, implicitly giving the time between each sample, as this is the inverted value. The choice of sampling rate in a digital coding of sound has not only a direct effect on the time resolution of the waveform, but it also determines the maximum possible frequency that can be represented. This is described in the Nyquist criteria, which states that a digital coding can only represent a waveform with frequencies inside the range of half the value of the sampling rate. The number of bits in a digital coding determines the binary numerical range in which the sampled amplitude values can be represented. Since the sample values always are a measurement of the magnitude of displacement relative to a maximum value, the numerical range is a representation of the amplitude resolution in the digital coding. The quality of digitally coded sound can be improved by increasing both the sample rate and the number of bits. The only trade-off is that these improvements require more processing and storing resources. A very common way of graphically displaying the sound waveforms is in time-amplitude plots generated from

the sampled values. These represent a tool that can be used for detailed time measurements of phonetic events in voice identifications.

## **Spectrum Analysis**

The traditional way of analysing the frequency domain of digitalised waveforms is through Fourier analysis. These are mathematical algorithms, which extract the frequency spectrum information of a sound from a set of samples. The set of samples are normally referred to as a window in the time domain, and increasing its size will improve the frequency resolution of the generated amplitude-frequency plot. The most commonly used of the available Fourier algorithms is the Fast Fourier Transform (FFT). As the name implies, this is because it is easy to implement in digital systems that requires a high performance. FFT is normally the algorithm by which digital spectrograms are generated. The usage of Fourier transforms in detailed formant frequency analysis of speech sounds has some limitations, caused by the complexity of the sound sources. When a window size that has sufficient frequency resolution is applied, the spectrum plot will have peaks for each harmonic frequency within the formant band, making it difficult to determine the actual resonating frequency causing the formant. To overcome this problem a specialised algorithm has been developed in order to model the properties of the articulation filter in the source-filter model of speech production. This method, which was first presented by Atal and Hanauer in 1971<sup>29</sup>, is called Linear Prediction Coefficients (LPC) and is used to estimate the mathematical transfer function of the filter. When an amplitude-frequency plot is generated from this transfer function, the peaks will represent the resonating frequencies of the vocal cavities, which correspond to the actual formant frequencies of a sound.

## **Parameters Selection**

One of the most important benefits of using techniques of automatic speaker recognition in forensic voice identification is that when parameters can be analysed individually, there is a possibility of applying a dynamic analysis procedure. Parameters can be chosen that are inside the range of available information in the recording of the unknown sample. If new techniques using new parameters are developed in the future, these can be added to the existing procedure expanding the range of possible choices. When one has the ability to analyse phonemes individually, the question immediately becomes: which parameter is best suited for voice identification purposes? As described earlier, intraspeaker variability is largely based upon variations in the user pattern of articulation and phonation. Parameters independent of this pattern would therefore theoretically be the ones with highest discrimination between speakers. In this context there are two types of phonemes that seem to be more promising than others. Some of the formant frequencies in nasal consonants are directly dependent on the nasal cavity, and since the size and shape of this is fixed, the speaker is unable to control these. High accuracy in automatic speaker recognition has been reported in systems solely based on these parameters<sup>30</sup>. The period of full closure of plosive consonants is also thought to be highly

independent of the habitual articulation pattern, as the duration of these are so short that the speaker cannot deliberately alter them. Other parameters that have been frequently used in automatic speaker recognition systems, although they are dependent on the habitual articulation and phonation pattern, is the formant frequencies of vowels<sup>31</sup> and the fundamental frequency in words and utterances<sup>32</sup>. Vowels are normally so long in duration during normal speech that the speaker has time to reach the static setting of articulation, making the formant structure very similar for each utterance. The fundamental frequency in speech sounds is altered during intonation of normal speech, but when its average value is measured over a period of time, it is known to settle at a value that is highly characteristic of a person. These features are just four examples of parameters frequently used in automatic speaker recognition systems. Many other parameters derived from both single phonemes and long time frequency spectra have been used, and the list is still expanding as more studies and experiments are conducted.

## **Parameter Performance**

From a forensic point of view, some of the most interesting studies on automatic speaker recognition are those which are conducted in order to statistically evaluate the performance of the different parameters involved<sup>33, 34, 35</sup>. In one of these, Sambur investigated a total of 92 parameters<sup>34</sup>. These included LPC analysis of formant frequencies in vowels and nasal consonants, LPC analysis of the frequency spectrum in fricatives, time measurements of both the closure period in plosives and the glides in diphthongs and the average fundamental frequency in words. As expected both formant frequencies in nasals and the duration of closure in plosives were some of the parameters that were found to have the highest discrimination between speakers. Other parameters that also were found to have a high performance included both formant frequencies in vowels and the average fundamental frequency in words. When a closed set experiment was performed with only the five parameters that were found to be the most discriminating (F2 in [n], F3 in [u], F2 in [i], duration of [k] and F3 in [m]) the predicted error rate of 0.03% was verified by an actual error rate of false identification of 0.003%. The findings in both this and similar studies cannot be directly applied to forensic voice identification, as the design of the experiments and the size of the reference population (usually about ten speakers) are not consistent with a forensic model. They do however indicate that high accuracy identifications and eliminations can be performed on a basis of only a few selected parameters and that it is possible to develop a statistical foundation to theoretically evaluate the performance of dynamic analysis.

## **Statistics of Voice Identification Evidence**

In using a dynamic analysis procedure, it is important in a forensic context to be able to estimate the statistical performance in each separate case. These can come in either the form of a likelihood ratio or as an estimated error rate. The mathematical methods used for calculations of error rates in the above-mentioned studies are all designed for the purposes of automatic speaker verification, and cannot be directly applied to forensic situations, as they do not account for all the encountered variables. Their statistical basis does however involve methods that have a potential of being adjusted to fit into a forensic context. I will in this chapter outline a basis of how statistics can be applied in forensic voice identification. This is based on some assumptions that need to be tested through further studies on the subject.

### **Data Base**

In calculating error rates or likelihood ratios, the distribution of the values of the parameters involved has to be analysed from a relatively large database of speakers. This database has to be specific for the language in question and should ideally be uniform regarding accents. It must also be sex dependent and represent a random selection of the population. Since it would be almost impossible to analyse each single phoneme in every possible phonetic context, the parameters have to be extracted from a random phonetic context. This implies that the database can also be used as a basis for text independent analysis, something that could be a big advantage in a forensic context. The values of each of the acoustic parameters of speech collected from this database are assumed to follow a normal distribution curve. This means that when the mean value and the standard deviation is known, the frequencies of specific values in a parameter can be calculated with the usage of developed statistical tools.

### **Error Rates**

A common way to statistically evaluate the relative performance of a parameter in automatic speaker recognition techniques is by applying the F-ratio. This is a measurement of the ratio of interspeaker to intraspeaker variations of a specific parameter, and is expressed mathematically as:

$$F = \frac{\frac{n}{m-1} \sum_{j=1}^m (\mu_j - \mu)^2}{\frac{1}{m(n-1)} \sum_{i=1}^n \sum_{j=1}^m (x_{ij} - \mu_j)^2}$$

Where  $\mu$  is the mean value of the  $m$  speakers in reference population,  $\mu_j$  is the mean value in the  $n$  reference sample from speaker  $j$  and  $x_{ij}$  is the  $i^{\text{th}}$  reference sample from

speaker  $j$ . When F-ratios are used to evaluate the general performance of parameters in a whole population, the parameters relative performance is increasing with their F-ratio values. They are therefore useful in order to select the most efficient subset of parameters used in a dynamic analysis. A common way of estimating error rates in automatic speaker recognition systems is to apply the *probability of error criterion*, which is derived from the F-ratios of the used parameters. This approach also seems promising for dynamic forensic analysis if the data for the reference population is collected from a sufficiently large database. It does however need to be adjusted in order to account for the possibility of false eliminations, something that has normally not been a subject of calculations in evaluations performed for automatic speaker recognition purposes.

## Likelihood Ratios

In order to account for intraspeaker variations in calculations of likelihood ratios, it is reasonable to say that a match of a specific parameter would only exist when the value in the unknown sample is within the range of the values extracted from the reference samples. When this is applied, the probability of observing the match will be equal to the area beneath the normal distribution curve restricted on each side by the interval of the reference samples. If a probability of match is in this way determined for all the parameters used in the analysis, the overall probability of observing the evidence given no-guilt will be the product of all the individual parameters. Calculating the likelihood ratio will then be a straightforward matter as the probability of observing the evidence given guilt is equal to one.

$$LR = \frac{p\langle E|C \rangle}{p\langle E|\bar{C} \rangle} = \frac{1}{p\langle match1 \rangle \times p\langle match2 \rangle \times \dots \times p\langle matchN \rangle}$$

If the value of a parameter in the unknown sample is outside the interval of intraspeaker variability in the reference samples, the probability that it in fact is not a match is dependent on the distance between the relative values and the general distribution of values in-between. The probability of observing this specific non-match will therefore be equal to the sum of areas beneath the normal distribution on each side of the interval between the unknown sample and the opposite border of the reference sample interval. If a probability of non-matches is in this way determined for all the parameters used in the analysis, the overall probability of observing the evidence given guilt will be the product of all the individual parameters. The calculation of the likelihood ratio can then be performed, as the probability of observing the evidence given no guilt in this situation is equal to one.

$$LR = \frac{p\langle E|C \rangle}{p\langle E|\bar{C} \rangle} = p\langle non-match1 \rangle \times p\langle non-match2 \rangle \times \dots \times p\langle non-matchN \rangle$$

In cases where the different parameters used in the analysis resolve in both matches and non-matches, the same procedure as above can be applied. The product of the probabilities of matches will still be the probability of observing the evidence given no-guilt, and the product of the non-matches are the probability of observing the evidence given guilt.

$$LP = \frac{p\langle E|C \rangle}{p\langle E|\bar{C} \rangle} = \frac{p\langle non-match1 \rangle \times p\langle non-match2 \rangle \times \dots \times p\langle non-matchN \rangle}{p\langle match1 \rangle \times p\langle match2 \rangle \times \dots \times p\langle matchN \rangle}$$

The assumptions this method is derived from includes not only that parameter values are normally distributed, but also that the ones used in the analysis are statistically independent. While this assumption might be reasonable for parameter sets such as duration of closure in plosives v. fundamental frequency, it might not be valid for parameters such as the duration of the closure in different plosives. It is therefore important to test every possible combination of parameters for statistical independence, and from this result develop a subset of parameters that can be used in the analysis.

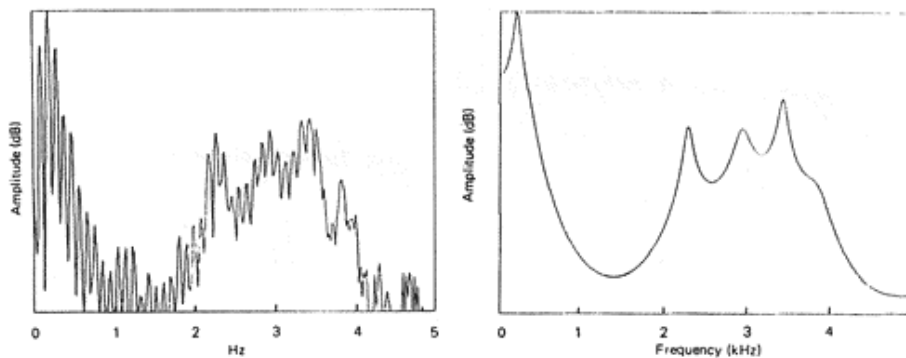


Figure 12: Fast Furrier Transform (FFT) and Linear Prediction Coefficient (LPC) frequency spectrum analysis of the vowel [i].

## **Conclusion**

The production of speech sounds is dependent upon three properties: the dimensions of the vocal organs, modes of phonation and manner of articulation. Every person has a habitual user pattern of phonation and articulation and it is individual differences in these, together with the shape and size of the resonating cavities, which make peoples voices sound different. The variations of the acoustic make up of speech sounds that are a result of these individual differences, are normally referred to as interspeaker variations. When one is analysing speech sounds for the purpose of voice identification, it is these variations one is looking for. Voice identification would be a relative simple task if people were more consistent in the way they speak. Intraspeaker variations are the term used to describe the differences in acoustic make up of speech sounds caused by the discrepancies in the usage of phonation and articulation. Both interspeaker and intraspeaker variations unfortunately affect most of the same parameters in speech sounds, something that makes voice identification a difficult task.

The traditional way of performing forensic voice identifications is by a combination of aural and spectrographic comparisons between the unknown sample and known reference samples. This procedure can be regarded as a subjective form of identification, as the examiner must the conclusions on the basis of his/her own judgement of the amount of similarity or dissimilarity between the spectrograms. In order to address this problem, several studies have been conducted on test examiners ability to make accurate conclusions regarding identity of speakers on the basis of spectrograms. These studies have reported huge differences in accuracy, as error rates have ranged from 0% to 83.3%. It has been argued that most of these inconsistencies may be a result of different designs of the experiments, and that the accuracy will increase with both the examiners experience and the possibility of making conclusions on the basis of confidence. The method of spectrographic voice identification has greatly evolved since it was first introduced, and is today controlled by standardized procedures. One survey on real forensic comparisons has shown that when these standards are applied, the technique can be utilised with a high degree of accuracy. The usage of spectrographic voice evidence is still an unsettled topic in the courts of law and different courts have ruled both for and against admissibility. The result of this forty-year-old battle is that spectrographic voice identification today, is a virtually dead field of forensic science. There is only a few law enforcement agenises and private investigators that still practice the technique on a regular basis.

One of the most promising prospects of future improvements of forensic voice identification has its basis in the huge amount of research performed on automatic speaker recognition. The different systems developed uses techniques and acoustic parameters that can be combined in order to develop a dynamic and objective procedure of forensic voice identification. This method may also overcome some of the problems encountered with admissibility of spectrographic evidence, as there exists possibilities of statistically estimating the accuracy for each separate case.

## Appendix A:

### Phonemes in Standard English

#### Vowels:

1. [i] as in **See** [si:]
2. [ɪ] as in **Sit** [sɪt]
3. [e] as in **Ten** [ten]
4. [æ] as in **Hat** [hæt]
5. [ɑ] as in **Arm** [ɑ:m]
6. [ɒ] as in **Got** [gɒt]
7. [ɔ] as in **Saw** [sɔ:]
8. [ʊ] as in **Put** [pʊt]
9. [u] as in **Too** [tu:]
10. [ʌ] as in **Cup** [kʌp]
11. [ɜ] as in **Fur** [fɜ:r]
12. [ə] as in **Ago** [ə'gəʊ]

#### Nasals:

1. [m] as in **Man** [mæn]
2. [n] as in **No** [nəʊ]
3. [ŋ] as in **Sing** [sɪŋ]

#### Plosives:

1. [p] as in **Pen** [pen]
2. [b] as in **Bad** [bæd]
3. [t] as in **Tea** [ti:]
4. [d] as in **Did** [dɪd]
5. [k] as in **Cat** [kæt]
6. [g] as in **Got** [gɒt]

#### Diphthongs:

1. [eɪ] as in **Page** [peɪdʒ]
2. [əʊ] as in **Home** [həʊm]
3. [aɪ] as in **Five** [faɪv]
4. [aʊ] as in **Now** [naʊ]
5. [ɔɪ] as in **Join** [dʒɔɪn]
6. [ɪə] as in **Near** [nɪə]
7. [eə] as in **Hair** [heə]
8. [ʊə] as in **Pure** [pjʊə]

#### Fricatives:

1. [tʃ] as in **Chin** [tʃɪn]
2. [dʒ] as in **June** [dʒu:n]
3. [f] as in **Fall** [fɔ:l]
4. [v] as in **Voice** [vɔɪs]
5. [θ] as in **Thin** [θɪn]
6. [ð] as in **Then** [ðen]
7. [s] as in **So** [səʊ]
8. [z] as in **Zoo** [zu:]
9. [ʃ] as in **She** [ʃi:]
10. [ʒ] as in **Vision** ['vɪʒn]
11. [h] as in **How** [haʊ]

#### Approximants:

1. [l] as in **Leg** [leg]
2. [r] as in **Red** [red]
3. [j] as in **Yes** [jes]
4. [w] as in **Wet** [wet]

## Appendix B:

### Reference Articles

- <sup>1</sup> W.Koenig, H.K.Dunn and L.Y.Lacy, *The Sound Spectrograph*, Journal of the Acoustical Society of America, vol. 18, pp19-49, 1946.
- <sup>2</sup> L.G.Kersta, *Voiceprint Identification*, Nature, vol. 196, pp 1253-1257, 1962.
- <sup>3</sup> R.H.Bolt et al, *Speaker Identification by Speech Spectrograms: A Scientists' View of its Reliability for legal purposes*, Journal of the Acoustical Society of America, vol. 47, pp 597-612, 1970.
- <sup>4</sup> M.A.Young and R.A.Campbell, *Effects of Context on Talker Identification*, Journal of the Acoustical Society of America, vol. 42, pp 1250-1254, 1967.
- <sup>5</sup> K.N.Stevens et al, *Speaker Authentication and Identification: A Comparison of Spectrographic and Auditory Presentation of Speech Material*, Journal of the Acoustical Society of America, vol. 44, pp 1596-1607, 1968.
- <sup>6</sup> O.Tosi, *Speaker Identification through Acoustic Spectrography*, Paper presented at XIV Int. Congress on Logopedics and Phoniatrics, Paris, Sept. 1968. (information from <sup>3</sup>)
- <sup>7</sup> O.Tosi et al, *Experiment on Voice Identification*, Journal of the Acoustical Society of America, vol. 51, pp 2030-2043, 1972.
- <sup>8</sup> R.H.Bolt et al, *Speaker identification by speech spectrograms: some further observations*, Journal of the Acoustical Society of America, vol. 54, pp 531-534, 1973.
- <sup>9</sup> J.W.Black et al, *Reply to "Speaker identification by speech spectrograms: some further observations"*, Journal of the Acoustical Society of America, vol. 54, pp 535-537, 1973.
- <sup>10</sup> B.Hazen, *Effects of differing phonetic contexts on spectrographic speaker identification*, Journal of the Acoustical Society of America, vol.54, pp 650-660, 1973.
- <sup>11</sup> A.R.Reich, *Effects of selected vocal disguises upon spectrographic speaker identification*, Journal of the Acoustical Society of America, vol. 60, pp 919-925, 1976.
- <sup>12</sup> National Research Council, *On the Theory and Practice of Voice Identification*, National Academy of Sciences, 1979.
- <sup>13</sup> Lt. Smrkovski, *Collaborative Study of Speaker Identification by the Voiceprint Method*, Journal of the Association of Official Analytical Chemists, vol. 58, pp 453, 1975.
- <sup>14</sup> B.E.Koenig, *Spectrographic voice identification: A forensic survey*, Journal of the Acoustical Society of America, vol.79, pp 2088-2090, 1986.
- <sup>15</sup> T.Shipp et al, *Some fundamental considerations regarding voice identification*, Journal of the Acoustical Society of America, vol. 82, pp 687-688, 1987.
- <sup>16</sup> B.E.Koenig et al, *Reply to "Some fundamental considerations regarding voice identification"*, Journal of the Acoustical Society of America, vol. 82, pp 688-689, 1987.
- <sup>17</sup> *Voice Comparison Standards*, Journal of Forensic Identification, vol. 41, pp 373-392, 1991.

- <sup>18</sup> *American Board of Recorded Evidence Voice Comparison Standards*, [www.aftiinc.com](http://www.aftiinc.com).
- <sup>19</sup> B.E.Koenig, *Spectrographic Voice Identification*, Crime Laboratory Digest, vol. 13, pp 105-118, 1986.
- <sup>20</sup> B.E.Koenig, *Selected Topics in Forensic Voice Identification*, Crime Laboratory Digest, vol. 20, pp78-81, 1993.
- <sup>21</sup> *People v. Straehle*, 12 NY L.F. 501, 1966.
- <sup>22</sup> *Frye v. US*, 54 App DC 46, 293 F 1013, 1923.
- <sup>23</sup> *US v. Addison*, 498 F.2d 741, 1974; *People v. Kelly*, 17 Cal. 3d 24, 1976; *People v. Tobey*, 401 Mich. 141, 1977; *People v. Collins*, 405 N.Y.S.2d 365,1978; *Reed v. State*, 391 A.2d 364, 1978; *Cornett v. State*, 450 N.E.2d 498,1983.
- <sup>24</sup> *Commonwealth v. Lykus*, 367 Mass. 191, 1975; *Hodo v. Superior Court*, 30 C.A.3d 778, 1973; *US v. Franks*, 511 F.2d 25, 1975; *People v. Rogers*, 385 N.Y.S.2d 228, 1976; *People v. Bein*, 453 N.Y.S.2d 343, 1982.
- <sup>25</sup> *Commonwealth v. Topa*, 471 Pa. 223, 1977.
- <sup>26</sup> *US v. Stifel*, 433 F.2d 431, 1971; *Commonwealth v. Lykus*, 367 Mass. 191, 1975.
- <sup>27</sup> *US v. Baller*, 519 F.2d 463, 1975; *People v. Rogers*, 86 Misc. 2d 868, 1976.
- <sup>28</sup> *State ex rel. Trimble v. Hedman*, 192 N.W.2d 432, 1971; *US v. Sample*, 378 F.Supp. 44, 1974.
- <sup>29</sup> B.S.Atal and L.Hanauer, *Speech Analysis and Synthesis by Linear Prediction of the Speech Wave*, Journal of the Acoustical Society of America, vol. 50, pp 637-655, 1971.
- <sup>30</sup> J.W.Glenn and N.Kleiner, *Speaker Identification Based on Nasal Phonation*, Journal of the Acoustical Society of America, vol. 43, pp 368-372, 1968.
- <sup>31</sup> M.J.Miles, *Speaker Recognition Based Upon an Analysis of Vowel Sounds and its Application to Forensic Work*, Thesis at University of Auckland New Zealand, 1989.
- <sup>32</sup> B.S.Atal, *Automatic Speaker Recognition Based on Pitch Contours*, Journal of the Acoustical Society of America, vol. 52, pp 1687-1697, 1972.
- <sup>33</sup> J.J.Wolf, *Efficient Acoustic Parameters for Speaker Recognition*, Journal of the Acoustical Society of America, vol. 51, pp 2044-2056, 1972.
- <sup>34</sup> M.R.Sambur, *Selection of Acoustic Features for Speaker Identification*, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 23, pp 176-182, 1975.
- <sup>35</sup> J.E.Atkinson, *Inter- and intraspeaker variability in fundamental voice frequency*, Journal of the Acoustical Society of America, vol. 60, pp 440-445, 1976.

## **Reference Books**

- J.Clark and C.Yallop, *An Introduction to Phonetics and Phonology*, Blackwell Publisher Ltd. 1995.
- M.J.Ball, *Phonetics for Speech Pathology*, Whurr Publisher Ltd. 1993.
- K.N.Stevens, *Acoustic Phonetics*, MIT press, 1998.
- R.K.Potter, G.A.Kopp and H.C.Green, *Visible Speech*, D.Van Nostrand Company Inc. 1947.
- National Research Council, *On the Theory and Practice of Voice Identification*, National Academy of Sciences, 1979.
- M.C.McDermont, T.Oven and F.McDermont, *Voice Identification: The Aural/Spectrographic Method*, [www.owlinvestigations.com](http://www.owlinvestigations.com).
- R.Saferstein, *Criminalistics an Introduction to Forensic Science*, Prentice Hall Inc. 1998.
- R.L.Klevans and R.D.Rodman, *Voice Recognition*, Artech House, 1997.
- S.J.Orfanidis, *Introduction to Signal Processing*, Prentice Hall International Inc. 1996.
- A.S.Hornby, *Oxford Advanced Learner's Dictionary of Current English*, Oxford University Press, 1974.
- C.J.Wild and G.A.F.Seber, *Chance Encounters*, John Wiley & Sons Inc. 2000.