# Automatic Speech Technologies in Forensic Phonetics. A term paper in Speech Technologies, GSLT 2007.

Håkan Burden
burden@chalmers.se
Dept. of Computer Science and Engineering
Gothenborg University and Chalmers University of Technology

February 24, 2008

**Abstract**

Automatic speech technologies can provide an objective and memory-constant analysis of how probable a suspect is as the perpertrator. Traditional techniques as speaker identification and verification can be used to determine the identity of a criminal. Speech technologies can also be used to create a profile of a perpertrator. Lie detecting claims to be a speech technology but there is no evidence that supports the claim.

## 1 Introduction.

In 1660 the English king Charles I was executed. Richard Gittens attended the execution and later testified that he had heard the executioner ask the king's forgiveness. Since the executioner's face was covered there were no eye witnesses who could identify the executioner. Gittens said that he had recognized the executioner as William Hulet by his speech. Hulet was accused and found guilty of high treason. He was sentenced to death.

Before the death sentence was carried out it was found that the execution had been performed by the ordinary hang man. As he confessed to the crime, Hulet was set free.

The situation arises in many cases even today. The villain is heard but not seen, there is an earwitness claiming to be able to indentify the perpertrator but what are the guarantees that the identification is correct? In Hulet's case the witness, Gittens, claimed he knew the identity of the executioner. Later evidence proved him wrong.

One can easily imagine a number of reasons for a jury to disregard or emphasize certain witnesses. The beheading of a king is a delicate business. It might have been that Gittens was right but the political consequences where more favourable towards sentencing the anonymous hang man. Or Gittens could have tried to exploit the situation and put Hulet in the frame. Or could it just be the case that Gittens had mistaken himself?

The account of Hulet's experiences of 17:th century justice is taken from Eriksson (2005).

## 2 Motivation.

Lies and deceptions are problems of old. Criminals use gloves to avoid leaving finger prints, they make sure to burn any left equipment to destroy DNA evidence and they disguise their voices when demanding ransom money by phone.

Using automatic speech analysis to identify the owner of a voice could be a step towards an objective identification of suspects. Just as fingerprints or DNA are considered to be unique, the voice is as well. Another benefit is that automatic speech technologies do not suffer from the deteriating effect of time which is the case for human memory.

For forensic purposes a set of different speech tests can also be performed and the combined result be used. Automatic speech technologies can also be used together with the testimony of a human earwitness, hopefully making the picture clearer.

The key issue for forensic phonetics is that voices are unique. Any implementation of speech technology to help solve problems regarding speech will need a way of digitally representing speech. The question then is how to do this and at the same time preserve the uniqeness of the voice.

When a digital representation of an analog signal is used to catch the nature of speech, the result will not only be a means of signal analysis but also a loss of information. As will be shown in section 4 a popular way of representing speech is by feature vectors. These speaker models not only capture the most important features of speakers. They can also be used to calculate the probability that the voice in a certain recording belongs to a suspect or to exclude a person from the suspects.

## 3 Historical background.

This section is based on Eriksson (2005).

In the late thirties work at the Bell Telephone Laboratories resulted in an invention called the Spectrograph. The official motivation for the project was that it could be used as an aid in pronounciation training for the deaf and students studying foreign languages. The project was despite this humanitarian motivation rated as a war project. Why would the American war industry put their money towards training equipment for the deaf during World War II? Probably because their real interest was in the possibility to develop a reliable speaker identifier. This could also explain why nothing was published until the end of the war (Grey and Kopp, 1944). The term voiceprint was first used in their report to refer to spectrograms.

In 1962 Kersta published a paper in Nature titled *Voiceprint identification* (Kersta, 1962). He claimed that correct identification could be obtained with 99% correctness by using voiceprints to compare two key words with each other. He was used as an expert witness by courts in a number of states.

His claim was not challenged until 1967. Young and Campbell obtained 78% correct identifications for training material consisting of two words spoken

in isolation by using voiceprints. If the same words were spoken in different contexts the number of correct identifications went down to 38% (Young, 1967).

Uptil now all work on speaker identification had been done by letting skilled phoneticians study the voiceprints in order to decide the similarity between the suspect and the perpertrator.

Since the early success of Kersta the use of voiceprinting has been questioned and today there is none in the speech science community that regards voiceprinting as a useful tool in forensic phonetics. Though there still are a number of private detectives who use the technique.

# 4 Speaker models.

In order to represent the speaker information in the analog speech signal some kind of digital speaker model is needed. The analysis of the speech flow is typically done in 20ms frames with 10ms overlap extracting a set of features to represent the speech information. The number of features usually lie around 14 to 16 and describe a vector of the same magnitude which is referred to as the speaker space.

The features cannot completely describe the speech production process and some information is lost. This problem is partly compensated for by how the models are built.

## 4.1 Mel-warped cepstra.

One of the more popular ways of extracting the features is by mel-warped cepstra.

The first step in obtaing a mel-warped cepstra is to apply fast Fourier transform to the speech signal in each window. This binds the signal to the frequency domain. In step two, Mel-warping is applied in order to give the lower bands more influence while diminishing the influence of the higher bands. Finally the logarithmic Mel-spectrum is passed on to the inverse fast Fourier transform which produces a cepstrum vector as output. The first samples, usually 14 or 16, are used as speaker features.

Voice production relies on pressing air between the vocal chords obtaining a periodic pulse that is then filtered through the mouth cavity into sounds. Cepstra represent the shape of the vocal tract. The properties making a voice unique reside in the periodic pulse. The pitch information can be omitted if the models are to only represent the speaker and not the spoken.

An important property of the cepstrum is that the overall spectral shape is modelled by the lower part of the cepstral vector and the harmonic structure is described by the higher order coefficients. Since the transfer function of the vocal tract is generally a smooth function, it will reside mostly in the lower part of the vector. Removing the pitch information is straighforwardly performed by omitting the higher order coefficients.

However, the most important justification of using cepstra is that they model human speech well. See Furui (1997), Reynolds et al. (2000) and Bimbot et al. (2004) for more technical details on obtaining models.

### 4.2 Possible shortcomings.

Returning to the question of the uniqeness of a voice a partial answer lies in the usage of features. Since the speaker space is finite, adding more speaker models into the speaker space will sooner or later mean that different voices will share speaker space. Ideally they are still distinguishable so that the digitalisation of the voice information into features does not hinder speaker recognition. However, this might not always be the case.

A more serious problem is speaker variability. Growing older and damage to the vocal tract has long-term effects on our voices. Stress, excitement, sore throats and blocked noses give temporary variations.

Unfamiliar voices are harder to remember and to distinguish from other voices for humans. A voice that has been heard over and over in a number of different circumstances is more familiar. The same goes for machines. There is no data like more data when training speaker models.

## 5 Speaker recognition.

The task of speaker recognition can be split into speaker identification and speaker verification. However, for forensic means the issue of speaker identification and verification can be the same as we will se later. A typical question for speaker verification is *Is the criminal suspect X?* while the question *Whom of the people that had access to the phone, made the call?* is answered by speaker identification. Automatic speech technologies can also be used to define a profile for the criminal.

Equal error rate, EER, is when the number of false acceptions equals the number of false rejections. This might do for some civil purposes but not in forensic analysis. Even if a high number of positive identifications is desirable it is even more important to keep the number of false identifications as close to zero as possible.

Even when speaker recognition fails automatic speech technologies can be useful to restore corrupted signals or creating profiles for the voice, by modeling such different aspects as sex, age, dialect, health, recording environment and channel.

While there can be big differences in the recording of a speaker depending on the speaker and the recording some issues like syntax, vocabulary and speech impediments can still be the same.

### 5.1 Open and closed sets.

This subsection relies on Furui (1997).

#### Open set.

If the task is to decide if the unknown speaker is someone in a group of people or not, the answer could either be one of the group members or a negative answer, if the voice is unknown. In order to solve the task a comparison with an absolute meaning has to be derived (see section 5.4).

**Closed set.**

The speaker to identify is within a set of known people (known in the sense that there is a model of their voices). The most probable candidate is then identified as the unknown speaker.

## 5.2 Text-dependency.

In forensic phonetics speaker recognition can be both text-dependent and text-independent.

**Text-dependent recognition.**

In text-dependent recognition both training and testing material is known. It can even be the same.

For instance, given a number of earlier recordings these can be used to obtain a corpus of utterances. A suspect can then be recorded while saying the utterances and the probability that these new utterances can be derived from a certain model be computed.

**Text-independent recognition.**

Text-independent recognition is the task of identification or verification when both training and test material can be any possible text.

When recording a new call, the model from the old recordings can be used to see how probable it is that the new call was done by the same caller.

In the context of text-independent speaker recognition a speaker produces a stream of features characterizing both the speech and the speaker. If the stream is longer than a few seconds the features are expected to capture speaker-dependent values instead of speech-dependent (Gish and Schmidt, 1994).

## 5.3 Speaker verification.

CIA has received a tape recording claimed to contain the comments of Usama bin Laden on the current US elections. Is the voice on the tape really bin Laden's?

Speaker verification is a special case of the open-set problem where the task is to decide if a certain speaker has the claimed identity.

For speaker verification in a text-dependent setting Hidden Markov Models are used. The state sequences then represent the allowed utterances. If the setting requires text-independent verification, Gaussian Mixture Models are often used. They use around 1000 components to represent the user utterances (Reynolds et al., 2000).

A model for every possible individual can be trained as described in section 4. If the recorded utterance $u$ from the speaker $x$ has a high probability of being given by the model of the said identity $i$ the claim is accepted, otherwise not

$$T \leq \frac{P(u|s_x)}{P(u|s_i)}$$

$P(u|s_x)$ is the probability that speaker $x$ said utterance $u$ and $P(u|s_i)$ is the same probability for the known identity $i$. If the quotient is larger or equal to the threshold $T$ the claim is accepted, otherwise rejected.

In phonetic forensics the threshold has to be high enough to not give any false accepts since this might mean that innocent people get convicted (Bimbot et al., 2004). Instead of using the threshold to accept or reject a claim it could be used as a confidence measure on how similiar the two models are.

## 5.4 Speaker identification.

An anonymous phone call demanding a ransom for the release of the owner of a nation-wide hi-fi chain has been recorded by the police. They now have a number of suspects who knew the whereabouts of the kidnapped. The police record all the suspects and compare their voices with that of the anonymous caller. The best case is that they get such a good match that they can arrest the kidnapper. The second best alternative would be to narrow down the number of suspects which might make the case easier to solve.

The likelihood that speaker model $i$ has generated utterance $x$ is given by Bayes law

$$P(i|x) = \frac{p(x|i)P(i)}{p(x)}$$

where $p(x|i)$ is the probability of speaker $i$ generating $x$ and $P(i)$ is the prior probability that the utterance came from speaker $i$, i.e. the probability given by other circumstances that speaker $i$ is guilty. Typically the prior probabilities are the same for all utterances and speakers but this does not need to be the case. The probability *p(x)* is the probability of $x$ being generated by any speaker

$$p(x) = \sum_{i=1}^{N} p(x|i)P(i)$$

where $N$ is the number of speaker models. It is worth noting that *p(x)* includes the probability for speaker *i*. Since *p(x)* is the same for all speakers and given that this also holds for $P(i)$ the speaker with the highest probability for the feature $x$ will be chosen. The importance of *p(x)* is that it normalizes the likelihood of speaker $i$ generating feature $x$ and therefore gives the resulting probability an absolute meaning.

The performance of the identification decreases as the number of models to work with grows (this section relies on Gish and Schmidt, 1994).

## 5.5 Splitting poor recordings into several models.

Telephones are one of the key instruments in recording voices committing criminal behaviour. Most threats are done by telephone and the tapping of a line is one of the more reliable ways of recording the coordination or dissemination of a crime. Telephone lines have limited bandwidth and frequencies below 300Hz are filtered out. This effects the quality of the recording.

The effects are not a result of the poor channel in its own. Not only can the channel result in loss of information and poor recording quality. It can also vary during the recording or from recording to recording if several telephone calls are used to obtain a model of the unknown speaker.

The number of calls made by cell phones or over the internet is growing. Due to the poorer quality of these channels the impact of the poor channel problem will be bigger in the near future.

One way of dealing with the problem is to define a speaker model for every recording or even splitting recordings of long enough length into several recordings, giving more speech models than recordings (Gish and Schmidt, 1994). The probability of an utterance given a suspect's speech model can then be compared to all other models.

By training several models from the same recording variations in channel, speaker quality and background noise etc. can be overcome. Each model will then be represent a subsequence of the utterance where certain qualities (like background noise) are constant. The speaker information in the a new recording can then be likewise broken down and compared to the obtained models.

The idea is that it is more informative to compare the utterance to a small model with constant background noice, channel distortion or speaker agitation than with a bigger model where the same parameters are varying. This bigger model has a more indefinite picture of the speaker and so comparing with it gives a vaguer idea of why there was a match.

It is more probable that a small model is an approximation of some of the information in the new recording than that all information in a recording is representative for all the information given by a large model.

Consider the case of 5 models for individual $i$ and 4 subutterances from suspect x. We can then fill in the squares in a 4x5 matrix with 1 if the verification was accepted. In this hypothetical case the threshold is set to 0.92.

$$\frac{P(u\,|s_x)}{P(u\,|s_i)} \geq 0.92 \Rightarrow 1$$

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | **i** | | | | | | |
| | | | | | | 1 | 1 | 1 | | |
| **x** | 1 | | | | | 1 | | | | |
| | | | 1 | | | | | 1 | 1 | |
| | | | | | | | 1 | | 1 | 1 |

In the left matrix only two matches were found which could be interpreted as a low probability that suspect $x$ is individual $i$. In the right matrix nearly half of the comparisons were accepted. Furthermore, if we know that in some of the rejected comparisons the model had a certain background noice or that the speaker sounded agitated and that this quality is not present in the particular subutterance, a rejection is not dismissive towards $x$ being $i$.

On the other hand, if both the model and the subutterance have the same agitation in voice but the verification claim was rejected it could be seen as evidence against a conviction.

## 5.6 Disguises.

In a report from 2000 Künzel (Künzel, 2000) claims that 15-25% of all cases dealt by the speaker identification section of the Bundeskriminalamt in Germany included some disguise of the voice. Even if low-level disguises like faked dialects or falsetto are easy to spot the effect on speaker recognition still leads to a major decrease in positive identifications, and are difficult to circumvent.

# 6 Lie detectors.

While automatic speech technologies can be used for speaker recognition and speaker profiling it cannot be used for detecting lies.

The most well-known lie detector is the Polygraph. It was first introduced in 1917 and has been somewhat refined since then. The idea behind the lie detector is that lying persons are under stress. Stress affects a person's breathing, pulse, blood pressure and can cause sweating. Stress will also alter the voice quality. The problem lies in interpreting the source of someone's stress as necessarily being lying. It could just as well be the effect of being interrogated with electrodes fastened to your skin while innocent.

There have been attempts to construct lie detectors using other voice and speech information. The idea is that the process of lying gives origin to micro tremors in the laryngal muscles which can be measured. The hypothesis has been tested and no such tremors could be found (Shipp and Izdebski, 1981).

Another attempt tries to establish that the passing of thoughts through the brain leaves a print on the speech flow (Nemesysco). It is not important what the interrogated says but how it is said. A lie will show up since the intention behind it will be spottable on the speech flow. There has been no findings of such links between speech and brain behaviour (Eriksson, 2005).

Still the business selling the detector has insurance companies as customers and also sell the same gadget as a love detector (Love detector).

# 7 Conclusions.

Speaker recognition relies on the notion of a voice being unique, just as a fingerprint. In order to analyze the analog flow of sound signals, feature vectors are used to get a digital representation of the speaker information. Thus the uniqueness of a voice can be lost or the more distinguishing marks be less marked. The problem is very small since the models are geared towards the lower order cepstral coefficients where the most information is.

The problem is instead adjusting to the varying recording qualities and the variations in the speaker's behaviour and health.

Instead of focusing on getting a conviction or ruling out a suspect there is the possibility to use the techniques from speaker verification to see how similiar the models of the anonymous caller are to the models of the known suspects. The threshold is then used as a confidence measure for every match instead of as a boolean result. What you get in the end is a two-dimensional matrix of confidence measures.

The big difference between automatic speech technologies and humans is that a computer is not affected by memory loss due to time in the same way as humans are. This also might make the computer more objective since its picture of things cannot be altered without new evidence being presented.

# References


F. Bimbot, J-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meigner, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacrétaz, and D. A. Reynolds. A turorial on text-independent speaker verification. *Journal on Applied Signal Processing*, (4):430–451, 2004.

Anders Eriksson. Tutorial on forensic speech science. In *Interspeech*, Lisbon, Portugal, 2005.

Sadaoki Furui. Recent advances in speaker recognition. *Pattern Recognition Letters*, pages 237–252, 1997.

Herbert Gish and Michael Schmidt. Text-independent speaker identification. *IEEE Signal Processing Magazine*, (October):18–32, 1994.

G. Grey and G. A. Kopp. Voiceprint identification. Technical report, Bell Telephone Laboratories, 1944.

L. G. Kersta. Voiceprint identification. *Nature*, (196):1253–1257, 1962.

H. Künzel. Effects of voice disguise on speaking fundamental frequency. *Forensic Linguistics*, (7):149–179, 2000.

Love detector. The love detector homepage. `http://www.love-detector.com`, 2008.

Nemesysco. The nemesysco homepage. `http://www.nemesysco.com`, 2008.

D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, (10):19–41, 2000.

T. Shipp and K. Izdebski. Current evidence for the existence of laryngeal macrotremor and microtremor. *Journal of Forensic Sciences*, (26):501–505, 1981.

M. A. Young. Effects of context on talker identification. *Journal of the Acoustical Society of America*, (42):1250–1254, 1967.