

HEALTH CARE UNIVERSITY COLLEGE GHEENT  
MEMBER OF GHEENT UNIVERSITY ASSOCIATION

# **TAPE AUTHENTICATION and VOICE IDENTIFICATION**

a case study in

# **FORENSIC ACOUSTIC PHONETICS**

Paul Corthals <sup>a,b</sup>  
John van Borsel <sup>b</sup>  
Kristiane van Lierde <sup>b</sup>

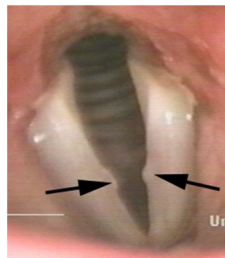
*August 24 2010*  
**OMICRON HALL**  
**11:30:00 AM**

**28<sup>th</sup> World Congress  
of the International Association  
of Logopedics and Phoniatrics**

<sup>a</sup> Faculty of Health Care  
Vesalius, University  
College Ghent, Belgium

<sup>b</sup> Faculty of Medicine  
and Health Sciences,  
Ghent University,  
Belgium

# *acoustic phonetics ~ speech & voice pathology*



**Fundamental  
Formants  
Intonation  
Prosody  
Accent  
Vocal folds  
Vocal tract  
Resonance  
Etc ...**

Paul Charhals me post 2010

We are all well acquainted with these concepts from voice pathology.

*acoustic phonetics ~ forensic applications*

**Fundamental  
Formants  
Intonation  
Prosody  
Accent  
Vocal folds  
Vocal tract  
Resonance  
Etc ...**



3

We are used to deal with these parameters, but only within the realm of voice pathology. We were asked by the local judiciary to examine a recording of a threatening telephone call. It was a case of speaker verification. There was this voice on the tape and there was a suspect. The question was: is the suspect the man on the tape? A second issue was to check the recording itself. We had to work on a copy of the original recording that was made in the victim's home. Although this copy was made by the local police, we had to demonstrate that our copy was not tampered with. In other words, we had to exclude the possibility that someone was trying to cheat us with a manipulated tape.

The only framework we had was... the framework of voice pathology, so we set out to apply it in the realm of forensics. In 1962, the term "voice print" was launched as a synonym for spectrogram, suggesting that the same level of certainty as in fingerprint analysis can be reached by spectrogram inspection. Obviously, that degree of certainty has never been scientifically confirmed. Indeed, compared to fingerprints, a speaker's speech signals are far more variable.

## *acoustic phonetics ~ forensic applications*

### Tape authentication

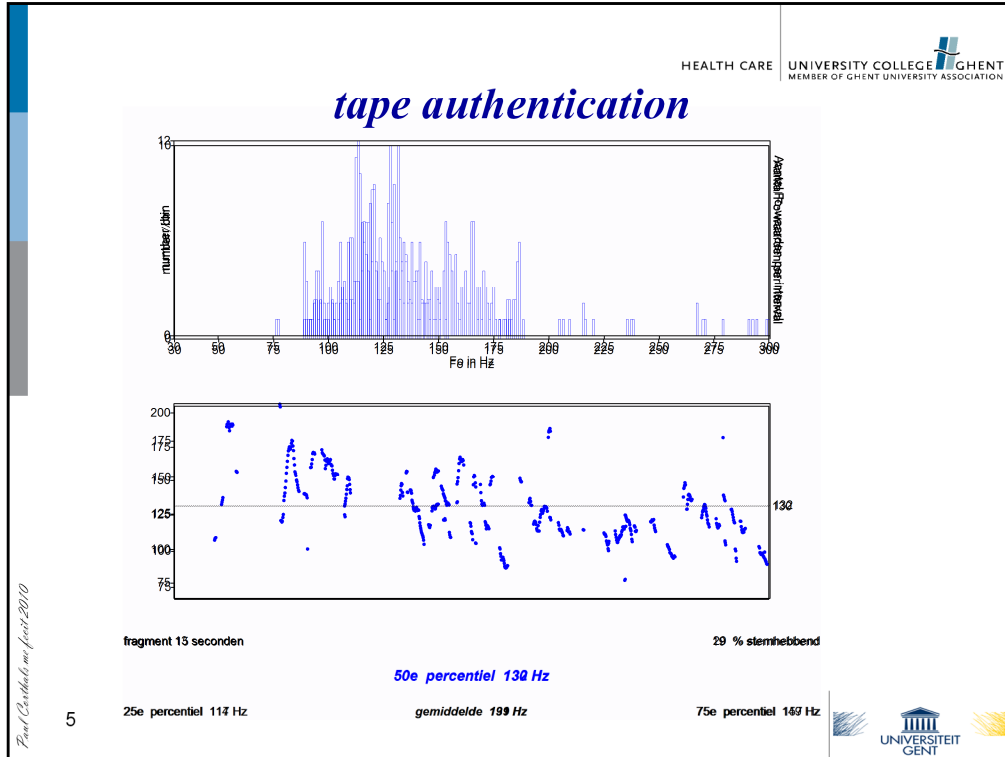
- was this tape tampered with to influence evidence?

### Voice identification

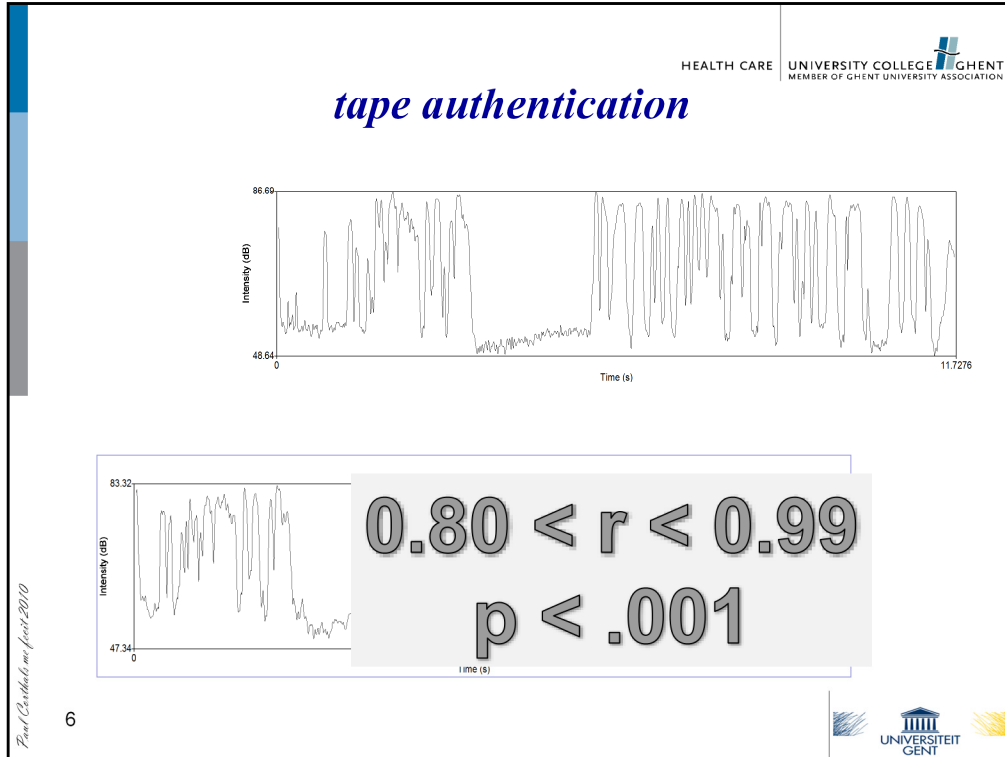
- is the recorded voice the defendants' voice?

Tape authentication is the examination of recordings in order to verify that their contents have not been edited and that they have not been manipulated otherwise in an attempt to influence judicial decisions. Tape authentication often implies other than purely phonetic techniques, e.g. sound engineering to detect non-speech evidence such as switching transients. However, acoustic-phonetic analysis of prosodic patterns such as intonation contours also is a feasible means to reveal discontinuities in a recording or to evaluate the congruence of two recordings. In our case, we had to work on a copy of the original exhibit recording.

Speaker recognition entails the attribution of a speech sample to a speaker using its acoustic-phonetic or perceptual properties as criteria. A widely used distinction is that between speaker verification versus speaker identification. In speaker verification, a speaker claims his own identity, for instance to gain access to privileged data or restricted areas. In contrast to speaker verification, speaker identification involves no identity claim by the speaker himself. Identification implies the selection of the author of a speech sample from an open or a closed set of possible speakers.



These are examples of pitch contours on the authors' copy and on the original tape. The dotted line is the 50<sup>th</sup> percentile of pitch values (130 Hz and 132 Hz respectively). So the median pitch on both recordings was essentially the same. The curves show instantaneous values of the voice fundamental of two equivalent stretches from both recordings. We calculated the correlation between all instantaneous pitch values.



The curves show instantaneous values of the voice intensity of two equivalent stretches from both recordings. We also calculated the correlation between all instantaneous intensity values. Pitch and intensity correlations turned out to be quite high and significant. The conclusion was that the copy we were about to use for speaker identification was not tampered with.

## *voice identification: voice lineup*

- 3 matched speakers (age, gender): “voice lineup”



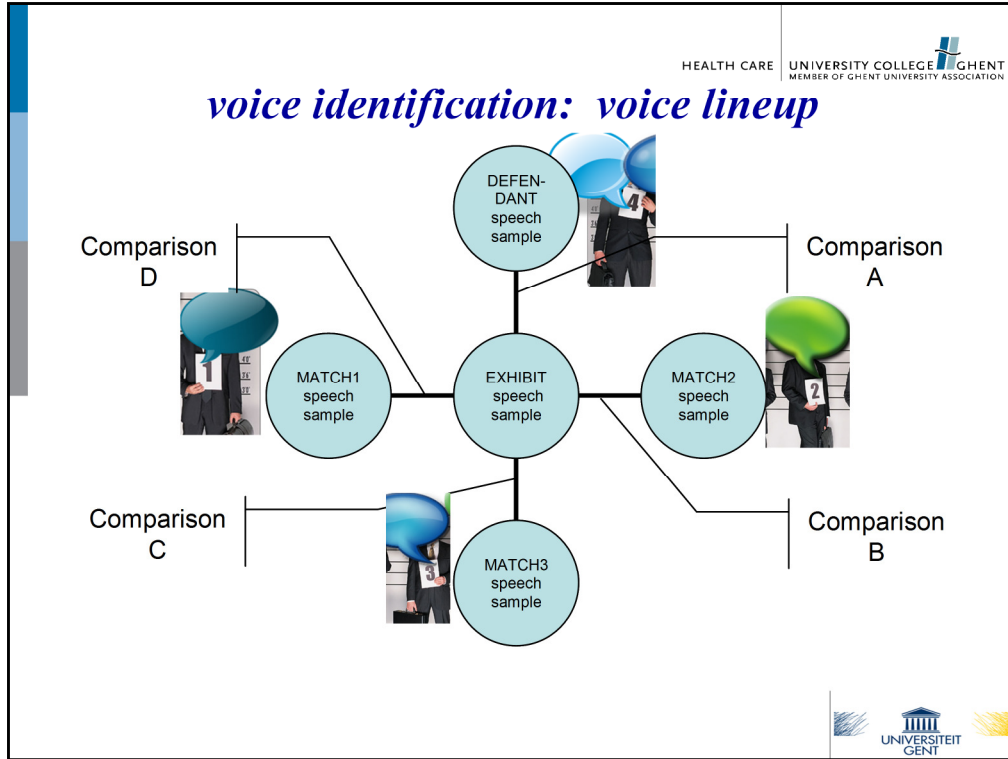
Paul Chartrand me, june 2010

7

Once the police copy of the exhibit tape was authenticated, we could address the main issue: speaker identification. You all know lineups from watching detective series. We know one of the men is the suspect and the other ones are matched persons. A witness is called in to compare their faces to the image he or she remembers.

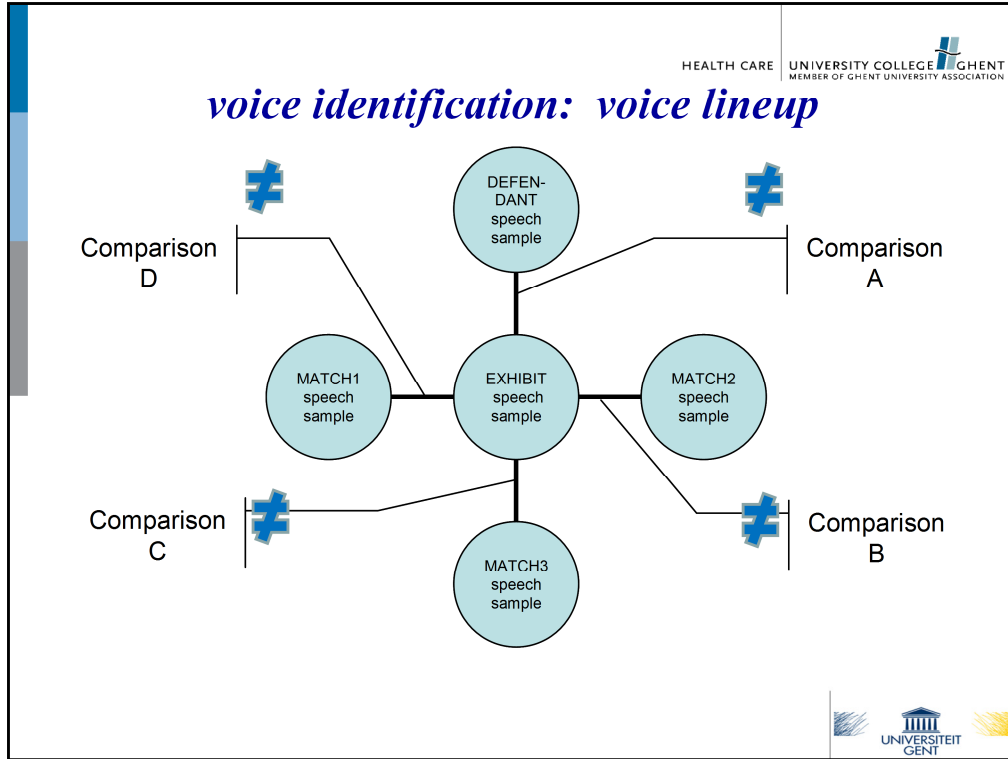
In perceptual speaker identification unbiased listeners compare recorded utterances. Such comparisons are hampered by the influence of the listeners' memory capacity (how long an utterance and how many voices can one remember accurately for comparison?), by the type of assignment (matching similar voices or identifying a speaker by selecting him/her from a so-called voice line-up, etc.) and even by the speaker ensemble, i.e. to whom exactly each speaker has to be compared (different ensembles may yield different perceptual identification results for a given target speaker). Also, more subtle psycholinguistic phenomena such as cue trading (the prominence of one acoustic trait compensating for the abstruseness of another one, resulting in an unchanged auditory perception) or verbal overshadowing (self-generated misinformation resulting from earlier attempts at verbal descriptions of a voice and making recognition abilities less accurate) can interfere in auditory judgements.

In a “technical” voice lineup the idea is to use a computer for comparing the voices of the defendant and the matched speakers with the voice that is on the exhibit tape. Of course, this comparison should be done by looking at speaker-specific acoustic features.

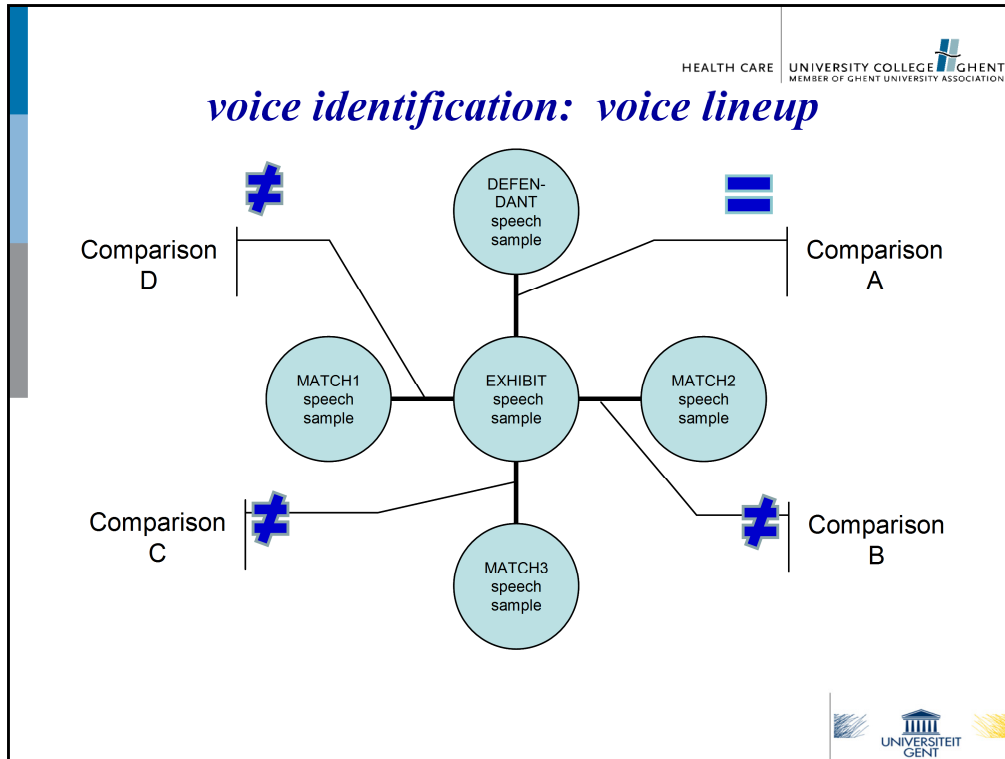


We compared the defendant's speech as well as the speech of three matched speakers with the exhibit speech sample. This makes four comparisons A,B,C,D.





Under the hypothesis that the defendant is not the speaker on the exhibit tape, we should find significant differences for all comparisons.



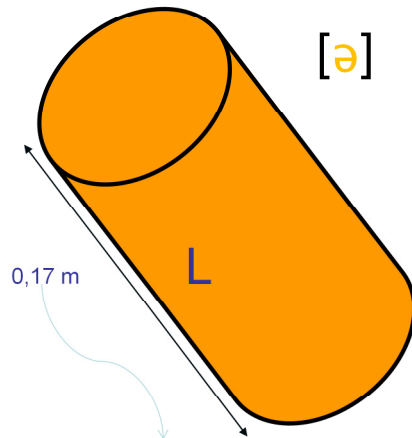
Under the hypothesis that the defendant is indeed the speaker on the exhibit tape, comparisons of type A should yield no significant differences whereas comparisons of type B, C or D should yield significant differences. We selected three acoustic markers that, according to the literature, have some potential as a speaker-specific feature. We used three acoustic markers, i.e. 3 features per comparison or per speaker. This is a total of  $3 \times 4 = 12$  comparisons with the exhibit voice.

## *voice identification: voice lineup*

- 3 matched speakers (age, gender): “voice lineup”
- reading assignment (words, some crucial)
- digital recording (\*.wav, 44100kHz sampling rate)
- 9 different vowels extracted (low, high, front, back)
- 63 data points per vowel for each feature
- 3 acoustic features
- **speaker-specific** features pertaining to
  - vocal fold** dimensions
  - vocal tract** dimensions

The search for stable speaker-specific acoustic details the search is still going on. Effective forensic speaker-dependent features due to laryngeal and vocal tract morphology emanate from parts or manoeuvres that show little variation during voicing and articulation and thereby introduce relatively invariant details.

## voice identification: formants



$$F_n = (2n-1)c/4L$$

where n is integer > 0, c is speed of sound in air

$$F_1 = (340\text{m/s}) / (4 \cdot 0,17) = 500 \text{ Hz}$$

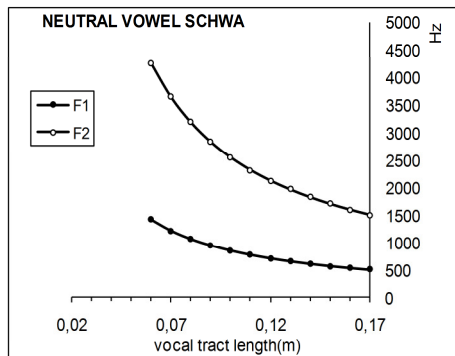
$$F_2 = 3 \cdot (340\text{m/s}) / (4 \cdot 0,17) = 1500 \text{ Hz}$$



The vocal tract resembles a hollow tube of 17 cm (male adult!), closed at the bottom by the vocal folds themselves. The vocal folds act like a drumstick, hitting the drum with air puffs. This kind of drum is more compatible with some frequencies than others, depending on its shape.

## voice identification: formants

[ə]



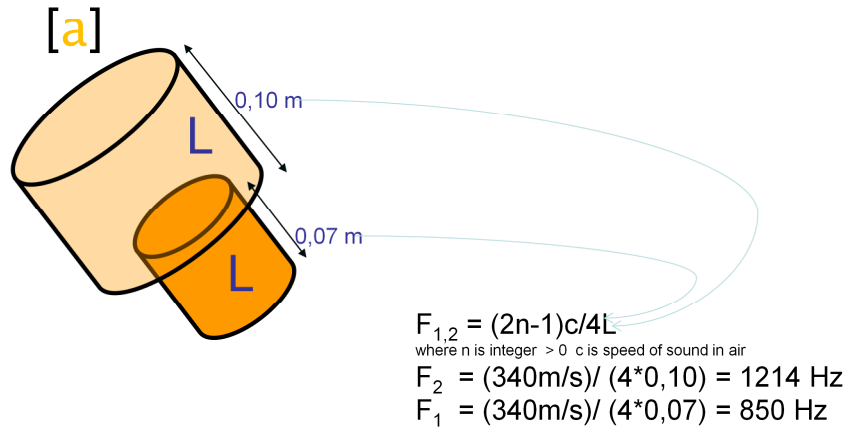
$$F_n = (2n-1)c/4L$$

where n is integer > 0, c is speed of sound in air

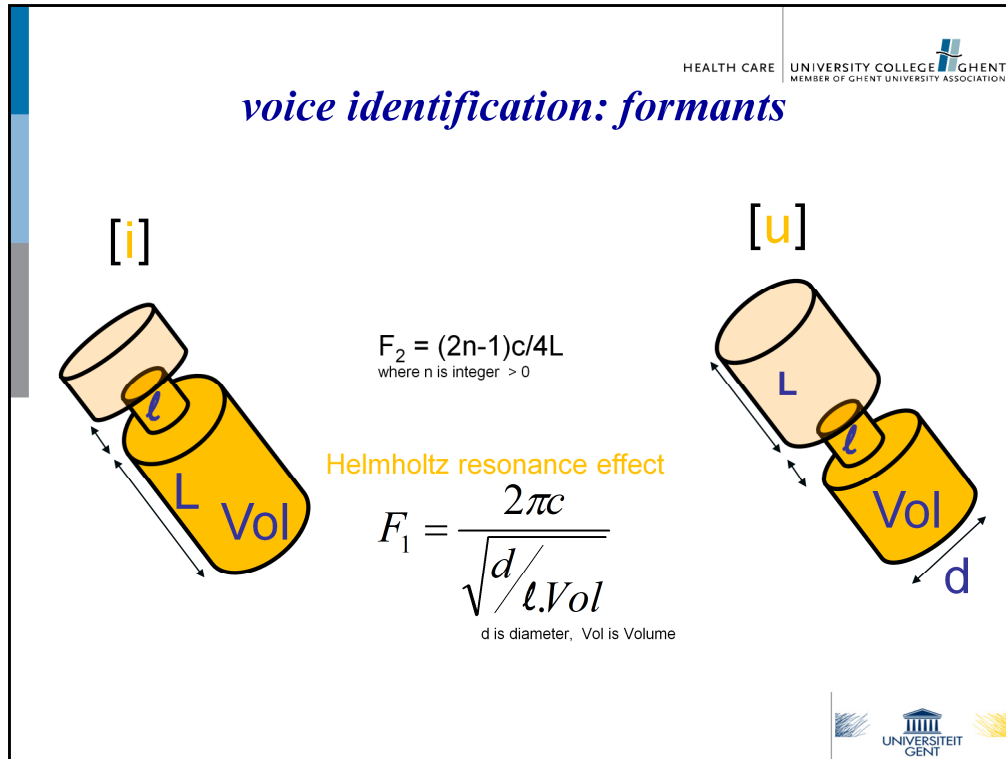


For the neutral vowel [ə] the value of L, vocal tract length, is the factor determining F1 and F2. For example: in adult males, the vocal tract is about 17 cm long. This results in 500Hz and 1500 Hz resonances. These are the formants of [ə]. If the vocal tract is shorter, both formants will be higher.

## voice identification: formants



For open vowels there are two tubes resonating in the same way. The upper tube corresponds to the oral part of the vocal tract and the lower tube corresponds to the pharyngeal part. The important thing to notice here is, again, that the formants depend on the dimensions of the vocal tract. Just inspect the equations predicting the formant values: L is the determining factor. This is why formants are good candidates as speaker-specific features.

*voice identification: formants*

For closed (or high) vowels we use a slightly different tube model. Above, we still see the “classic” tube, but now there is a bottle-shaped Helmholtz resonator below it. The bottleneck corresponds to the tongue constriction and the bottle corresponds to the pharynx. Its resonance effect can be calculated with an appropriate equation. Capital letter  $L$  stands for the size of the bottle.  $V$  stands for the volume of the bottle. Small letter  $l$  stands for the length of the bottleneck. The Helmholtz resonance is quite low (maybe you remember that from that game we all played: blowing over an empty bottle to make that low spooky sound...). The bottle will deliver the first formant. The second formant is coming from the lower tube for front vowels and from the top tube for back vowels. Again we can say that the dimensions of the vocal tract (the bottle and the resonating tube on top of it) are proportional to the formants. In that sense, they are speaker-specific features.

The position of the bottleneck corresponds to the tongue placement. It plays a role in the values of  $L$  and  $V$ . We could say that variations in the bottleneck position for a particular vowel, reflect the tongue placement habits of the speaker. Again: this is a speaker-specific feature.

*voice identification: pitch*

$$\text{frequency} \sim \frac{\text{stiffness}}{\text{mass}}$$

The vocal folds are so-called free oscillators. This means there is no time-dependent force controlling the rhythm of their movement. Free oscillators vibrate at their natural frequency. The natural frequency of the vocal folds is the voice fundamental. It determines voice pitch. The natural frequency (in other words: voice pitch) is determined by stiffness and mass of the vocal folds. The larger their mass, the lower their frequency. This is why children have a higher voice than adults and this is why boys get a lower voice in puberty. This mass is speaker-specific, stiffness is not since a speaker can change the tension of his vocal folds. Therefore, voice pitch can be considered a speaker-specific feature, but only to a certain extent.





## voice identification: pitch

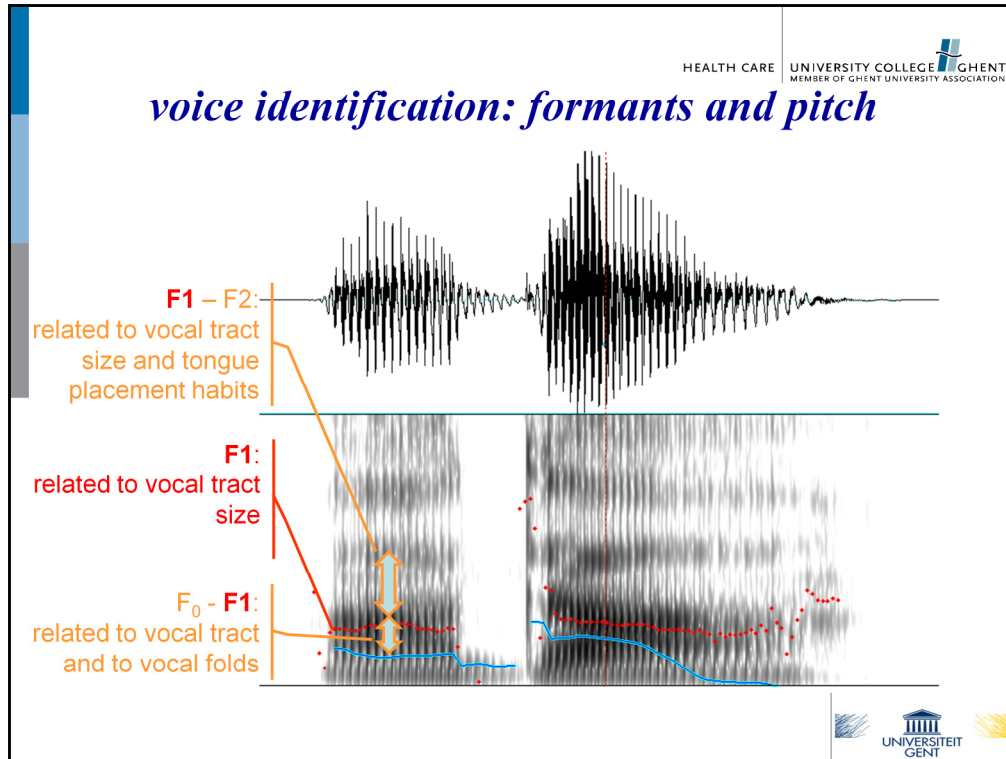
$$F_n = \frac{\text{constants}}{L}$$

$$F_1 = \frac{\text{constants}}{\sqrt{d/l \cdot Vol}}$$

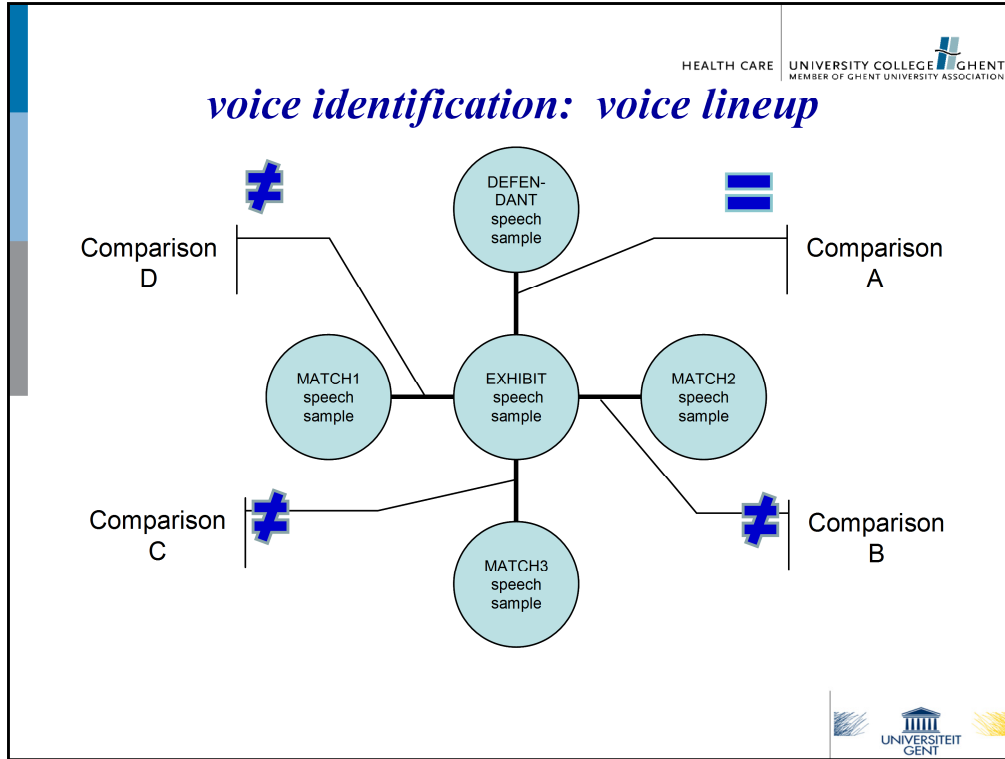
$$F_0 \sim \frac{\text{stiffness}}{\text{mass}}$$

 constants  
 speaker-specific

To sum up, the acoustic features we used were formants and voice fundamental. If we disregard the constants in the equation, we see mostly speaker-specific parameters.



The three features we extracted revolve around the first formant. The first formant is a good candidate for the list of speaker-specific features. On top of that, it always carries the bulk of the energy in the signal, which makes it easy to discern. The other features are combinations of other speaker-specific information and the first formant. For a given vowel, the F1-F2 distance combines the speaker specificity of F1 and F2 (the overall size of the vocal tract) and the speaker's habits in tongue placement. The F0-F1 distance reflects the overall size of the vocal tract and the size of the vocal folds (but not their tension).



Remember that under the hypothesis that the defendant is indeed the speaker on the exhibit tape, comparisons of type A should yield *no* significant differences whereas comparisons of type B, C or D *should* yield significant differences.

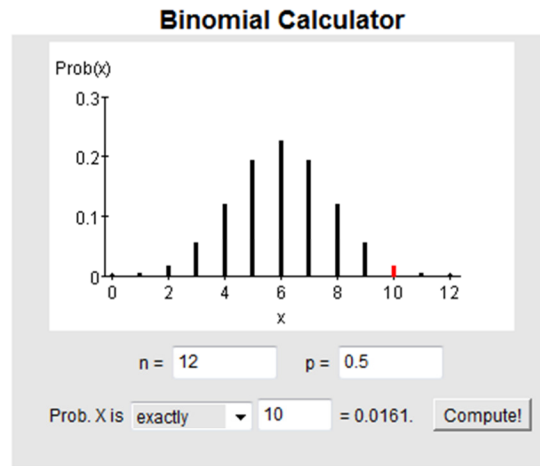
*voice identification: voice lineup*

Feature	Comparison A (defendant- exhibit)	Comparison B (speaker1- exhibit)	Comparison C (speaker2- exhibit)	Comparison D (speaker3- exhibit)
Absolute value 1 <sup>st</sup> formant	<i>NO significant difference</i> $p=0.300^{**}$	<i>Significant difference</i> $p<0.001^{**}$	<i>Significant difference</i> $p<0.001^{**}$	<i>Significant difference</i> $p<0.005^{**}$
Relative position 1 <sup>st</sup> formant and fundamental	<i>NO significant difference</i> $p=0.851^{**}$	<i>Significant difference</i> $p<0.001^{**}$	<i>Significant difference</i> $p<0.001^{**}$	<i>Significant difference</i> $p<0.005^{**}$
Relative position 1 <sup>st</sup> formant and 2 <sup>nd</sup> formant	<i>NO significant difference</i> $p=0.131^{**}$	<i>Significant difference</i> $p<0.001^{**}$	<i>NO significant difference</i> $p=0.112$	<i>NO significant difference</i> $p=0.076$

hit

miss

## *voice identification: voice lineup*



chances of obtaining 10 hits in 12 trials

If the chance of making a hit by chance is 50%, then the most frequent outcome is 6 hits out of 12 trials. The chance of obtaining 10 hits in 12 trials is less than 5% (the exact binomial probability is 1.6%).

We chose the binomial statistical model to calculate what random outcomes we would obtain if only chance was in play. One could argue that the 12 trials were not entirely independent.

## *Tape authentication & voice identification conclusions*

- voice pathology concepts (pitch, formant structure) can be seen as speaker-specific features for forensic applications
- forensic phonetics has no definitive set of target features and no universally accepted protocol to underpin identification or authentication procedures, experts have to adapt their methods to the available material in each case
- forensic phonetics is not the only element in a judge's decision...

HEALTH CARE UNIVERSITY COLLEGE GHEENT  
MEMBER OF GHEENT UNIVERSITY ASSOCIATION

NIEUWS 9

acquittal in spite of overwhelming evidence

Commissaris vrijuit ondanks verpletterende bewijslast

is vrijgesproken van stalking en bedreiging

Judge: "... while the voice lineup gives enough *statistical* certainty... that is not enough for this court"

was zelf nooit betrokken partij in de bewuste tuchtdossiers. En toch kreeg hij die vreemde telefoontjes. Het onderzoek bracht hem had bedreigd. Maar de rechter besloot toch om de weinige twijfel die er nog is in het voordeel van de verdachte te laten pleiten. Het is niet boven alle redelijke twijfel vast dat hij de dader is', voorzitter.

'In de loop van het onderzoek werden ook nog andere personen aangeduid die niet helemaal uitgesloten kunnen worden. Ook is er geen motief. Het dreigement toont geen verband met de zaak aan. Van de drie getuigen werd wel aangetoond dat ook anderen daar toegang toe hadden, bijvoorbeeld in de kantine. Tot slot wijst de stemtest op een voldoende statistische zekerheid, maar dat is onvoldoende voor deze rechtbank.'

'Dit is onbegrijpelijk: alles wijst in de richting van de verdachte. Dit is onbegrijpelijk. Alle elementen van het onderzoek wezen in zijn richting. Nochtans heb ik hem zelf nooit als verdachte aangeduid.'

'En toch heeft deze zaak mijn leven draaischijf veranderd: ik ben veranderd van job, genoeg al mijn ex-collega's haten mij en ook voor mijn gezin is dit heel zwaar geweest. Er werd ook druk op mij uitgeoefend om te stoppen met deze rechtszaak. En dit is dan het resultaat. Ik overweeg om in beroep te gaan.'

jaar van 2003 tot drie keer toe een dreigefoonie. Daarin werd hij met de dood bedreigd 'als hij niet met zijn poten van zijn wijven zou blijven'. Maschelein was totaal verrast en diende een klacht in tegen onbekenden. Nooit had hij kunnen denken gesteld. Daarom ook denken we dat deze zaak een afrekening is van de mensen die toen ter verantwoording werden geroepen', zegt zijn advocaat Paul Oudejans. Maar Maschelein van

FREDERIK VELDRE

Paul Cavaliere mei juni 2010

23

UNIVERSITEIT GENT

This is what we read a few months later in the local newspaper. In forensic phonetics, identification techniques fail to provide absolute certainty in that the output most often is a probability statement expressing the chances of a particular result occurring or a statistical type II error. The fundamental problem is that statistical significance is not accepted in court, because of the margin of uncertainty, however small that may be. Indeed, statistical certainty always approaches but never equals 100%.