

Visual Acoustic vs. Aural Perceptual Speaker Identification in a Closed Set of Disguised Voices

Jonas Lindh

Department of Linguistics
Göteborg University

Abstract

Many studies of automatic speaker recognition have investigated which parameters that perform best. This paper presents an experiment where graphic representations of LTAS (Long Time Average Spectrum) were used to identify speakers from a closed set of disguised voices and determine how well the graphic method performed compared to an aural approach.

Nine different speakers were recorded uttering a fake threat. The speakers used different disguises such as dialect, accent, whisper, falsetto etc. and the verbatim "threat" in a normal voice.

Using high quality recordings, visual comparison of the Praat "vocal tract" graphs of LTAS outperformed the aural approach in identifying the disguised voices. Performing speaker identification aurally does not mean analyzing a different sample than the one being analyzed acoustically. Studies of aural perception show a hypothesizing, top-down, active process, which create interesting questions regarding aural speaker identification with bad quality recordings in noisy backgrounds etc. However, more tests on telephone quality recordings, authentic material and additional types of acoustic measurements, must be performed to be able to say anything about LTAS with implications for forensic purposes.

Background and Introduction

The so-called "voiceprint" approach introduced by Lawrence Kersta (1962) suggested a pattern matching procedure comparing broadband spectrograms for speaker identification purposes. It is within this context that an interest in studying visual vs. aural methods arose. Since complex visual pattern matching activates the right hemisphere of the brain and speech- and language processes usually the left (Rose, 2002) it would be preferable to find a way to integrate both. There are many problems to be considered when using visual representations of acoustic data within the context of forensic speaker identification, especially considering

the effects of low quality recordings. Generally, one can say that primarily aural identification has been the leading method when it comes to casework. Many studies have been carried out to see what parameters are most stable or where effects of low quality can be calculated, for example the telephone effect (Künzel, 2001).

Generally, LTAS becomes rather stable after 30-40 seconds of speech. (Boves, 1984; Fritzell et. al., 1974; Keller, 2004) LTAS reflects the energy highs and lows generated by the vocal tract filter on average, which means that it should be more difficult to alter than, for example, F0 or specific phones, why this measure is often chosen to visually represent the general energy distributions in frequency for the speech signal. Several studies have been conducted to study energy ratios and level differences for LTAS (Löfqvist, 1986; Löfqvist & Manderson, 1987; Gauffin & Sundberg, 1977; Kitzing, 1986). Kitzing (1986) recommended that patients should read at the same degree of vocal loudness to avoid the differences that occurred especially in higher frequencies. Kitzing & Åkerlund (1993) pointed out the need for an investigation of the effect of vocal loudness on LTAS curves. Nordenberg & Sundberg (2003) performed such a test and showed that vocal loudness and varied f0 gave variations in Long Time Average Spectra. However, even though an expected variation has been shown, the ability to perform pattern matching on the graphs seems to be possible. It has been observed that a slight difference between the identification results between subjects depends on whether they consider distance more important than shape/pattern. Hollien & Majewski (1977) tested long-term spectra as a means of speaker identification under three different speaking conditions, i.e. normal, during stress and disguised speech. LTS for fifty American and fifty polish male speakers were used under fullband as well as passband conditions. The results demonstrated high levels of correct identification (especially under fullband conditions) for normal speech with degrading results for stress and disguise.

Method

The sixteen disguised voices and “suspects” (references), were recorded by six females and three males. The recordings were made with a high quality microphone in front of a personal computer and the subjects recorded one “normal” and as many disguised voices as they wanted, repeating the same fake threat in Swedish. All recordings were between four and six seconds long and sampled at 16kHz. Forced choice was applied in both the aural and visual tests.

The Graphic Representations of LTAS

The “vocal tract” function in Praat draws the LTAS envelope (in decibel) as if they were vocal tract areas (in square meters). This gives a graph representing the LTAS. The graph does not give the axis values, which is reasonable since the overall absolute amplitude, as a parameter, has no real value (Nordenberg & Sundberg, 2003). The important information lies in the relative spectral envelopes represented by the line showing the energy distribution as a function of frequency.

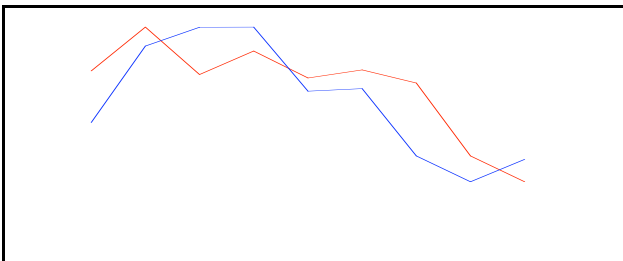


Figure 1. A graph comparison sample (in the test the target is red and each reference blue).

The graphic representations of LTAS were created from an LTAS object using 100 Hz frequency bins. (Boersma & Weenink, 2005)

The Visual Comparison Test

Graphs representing LTAS were created for sixteen disguised voices and paired up with each of the reference samples to be used in a visual identification test performed by ten subjects. The order in which they were presented was randomized. The subjects were all students or employees at the Department of Linguistics, Göteborg University. They had all, at some point, taken an undergraduate course in phonetics and/or speech technology.

The subjects compared each disguised voice with all the suspects/references in pairs and

then decided which one they thought was the most similar one comparing both shape and/or distance. The subjects were also told that the graphs had no timeline and that they were supposed to perform pattern matching, answering which graphs were the most similar ones in each test sample. They were also asked to comment on how they reached each conclusion and if distance or shape was most important when coming to a decision. This was done to be able to interpret how subjects compared the visual input. They were allowed to inspect the graphs as many times and as long as they wanted.

The Aural Identification Test

Seven subjects performed aural identification on the same set of samples to be able to compare the results easily.

The recordings were put in a list in a randomized order. Subjects used headphones and could listen to the samples as many times as they wanted before deciding which one of the references they thought sounded most like the target. All subjects were of the same category as in the visual test. Some test subjects were the same as in the visual test.

Results and Discussion

Even though there is a great difference in performance between subjects within each test, it is clear that the visual identification outperforms the aural.

The Visual Identification Results

The results for the visual tests show consistency.

Table 1. Inter-rater Reliability Analysis (Cronbach's alpha).

N of Disguised Voices	16
N of Subjects	10
Alpha	0.91

The impression based on the comments is that subjects with a preference for pattern and shape rather than distance generally performed better in the visual test.

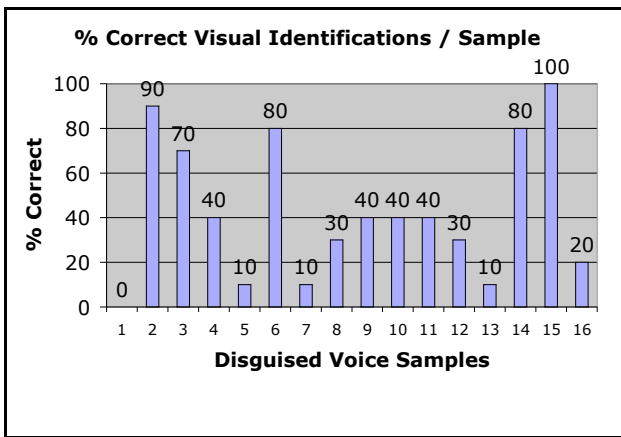


Figure 2. Percent correct visual identifications per sample (16) for 10 subjects.

Figure 2 shows how many correct identifications that were made per disguised voice sample. Some graphs were obviously very difficult to identify. Why that is so, or how those graphs differ, has not yet been investigated.

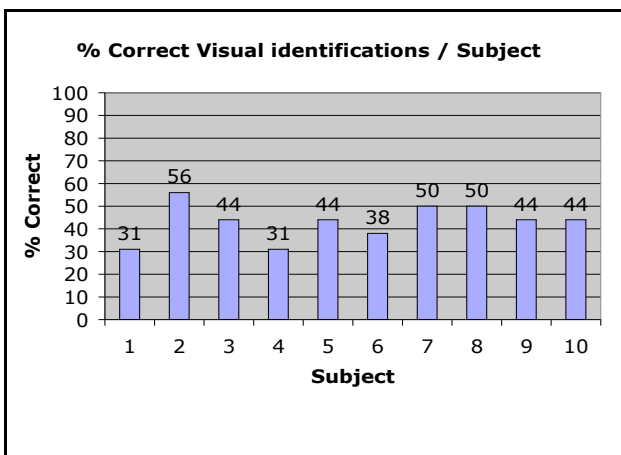


Figure 3. Percent correct visual identifications per subject (10) for 16 samples.

Figure 3 shows the identification results for each subject, which varies from nine correct identifications to five. As mentioned above the performance was clearly related to whether the subject used pattern/shape matching more than distance. The average identification score for the visual test is 6.9, which could be considered as rather low, but considering the difficulties presented in the aural test results it is merely the comparison which is taken into consideration in this study.

The Aural Identification Results

The results in the aural test were less correlated. The reason is simply that subjects found the task much more difficult, i.e. most subjects

thought that “no decision” should have been added as an alternative answer.

Table 2. Inter-rater Reliability Analysis (Cronbach's alpha).

N of Disguised Voices	16
N of Subjects	7
Alpha	0.83

The reliability score is lower in this test compared to the visual test. However, the correlation is high enough to be interpreted as a rather high correlation between subjects.

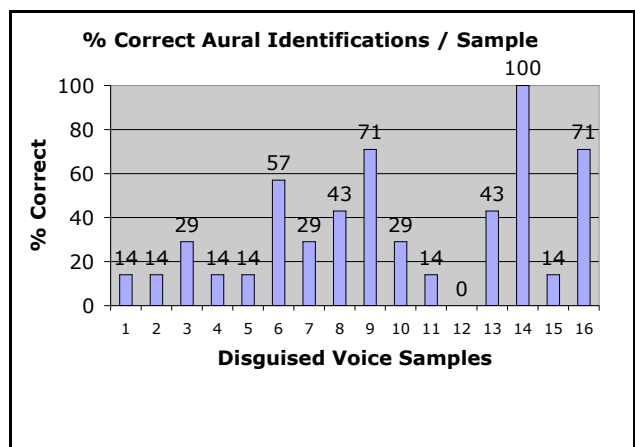


Figure 4. Percent correct aural identifications per sample (16) for 7 subjects.

Figure 4 gives a result overview, which may be compared with the corresponding Figure 2 for the visual test. The amount of correct identifications per sample is significantly lower though the maximum is lower (seven subjects vs. ten).

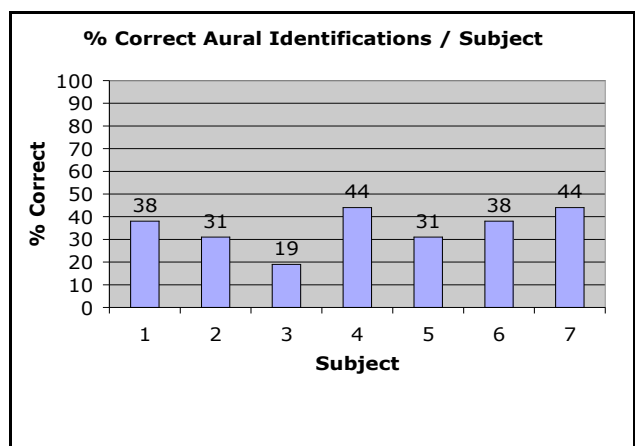


Figure 5. Percent correct aural identifications per subject (7) for 16 samples.

Figure 5 presents the figures corresponding to table three in the visual identification test. The subjects' results are significantly lower even though lowest visual score (five) is higher than the highest aural score (seven). Since there seems to be an individual strategy success involved. The result per subject in the aural test also shows a higher degree of variation than the visual. This is probably due to the difficulties they showed in deciding on which reference to choose.

Conclusions

General advantages with graphic representations are:

- Intra subjectively applicable (depending on the amount of data).
- Relatively simple fundamentals for calculation.
- Rather easy to visualize.

General disadvantages are:

- Difficult to quantify and substantiate comparisons.
- The visualization depends on F0 and vocal loudness variations.
- An average always ignores specific events in the speech signal.

Considering the categorical, top-down active human speech perception process (Grosjean, 1980), it is interesting to find complementary visual acoustic information to aural methods in forensic speaker identification. When two voice samples are compared, the same input is judged no matter if it is aurally or acoustically. The question is how it is analyzed and how the acoustic visual and the aural perceptual information are processed. If a better understanding between the two is reached, objective methods can be used to judge similarities. Objective acoustic methods can also more easily be excluded on well-grounded arguments as well as subjective aural ones. This could also lead to better statistical data in forensic speaker identification if computer based methods can be used with more confident supervision. It is clear that aural mistakes are made, especially for disguised voices.

The graphic representations used in this experiment are not claimed to be complete images reflecting the voice of a speaker. They are but examples showing that in some cases visual acoustic input are better at discriminating between speakers than are ears alone.

References

- Boersma P. & Weenink D. (2005) Praat: doing phonetics by computer (Version 4.3.01) [Computer Program]. Retrieved from <<http://www.praat.org/>>
- Boves L. (1984) The phonetic basis of perceptual ratings of running speech Foris Publications, Dordrecht.
- Gauffin J. & Sundberg J. (1977) Clinical application of acoustic voice analysis. Part II: Acoustic analysis, results 1977/2-3: 39-43.
- Grosjean F. (1980). Spoken word recognition processes and the gating paradigm. Perception and Psychophysics, 28, 267-283.
- Hollien H. & Majewski W. (1977) Speaker identification by long-term spectra under normal and distorted speech conditions. Journal of the Acoustical Society of America 62: 975-980.
- Keller E. (2004) The analysis of voice quality in speech processing. In Lecture notes in computer science, Springer Verlag, Berlin.
- Kersta L. G. (1962) Voiceprint identification. Nature 196: 1253-1257.
- Kitzing P. (1986) LTAS criteria pertinent to the measurement of voice quality. Journal of Phonetics, 14: 477-482.
- Künzel H. J. (2001) Beware of the 'telephone effect': The influence of telephone transmission on the measurement of formant frequencies. Forensic Linguistics 8: 80-99.
- Löfqvist A. (1986) The long-time-average spectrum as a tool in voice research. Journal of Phonetics, 14: 471-475.
- Löfqvist A. & Mandersson B. (1987) Long-time average spectrum of speech and voice analysis. Folia phoniatica, 39: 221-229.
- Nordenberg M. & Sundberg J. (2003) Effect on LTAS of vocal loudness variation. In: TMH-QPSR, KTH, 45: 93-100.
- Rose P. (2002) Forensic Speaker Identification. New York, Taylor & Francis.
- Stevens K. N. (1993) Lexical access from features. In Speech communication group working papers (Vol. VIII, p. 119-144). Research Laboratory of Electronics, Massachusetts Institute of Technology.