A comparison between human perception and a speaker verification system score of a voice imitation

Elisabeth Zetterholm[1], **Mats Blomberg**[2], **Daniel Elenius**[2]

[1]Department of Philosophy & Linguistics, Umeå University, Sweden
[2]Department of Speech, Music and Hearing, KTH, Stockholm, Sweden

Abstract

A professional impersonator has been studied when training his voice to mimic two target speakers. A three-fold investigation has been conducted; a computer-based speaker verification system was used, phonetic-acoustic measurements were made and a perception test was conducted. Our idea behind using this type of system is to measure how close to the target voice a professional impersonation might be able to reach and to relate this to phonetic-acoustic analyses of the mimic speech and human perception. The significantly increased verification scores and the phonetic-acoustic analyses show that the impersonator really changes his natural voice and speech in his imitations. The results of the perception test show that there is no, or only a small, correlation between the verification system and the listeners when estimating the voice imitations and how close they are to one of the target speakers.

1. Introduction

Imitation often sounds convincing. For several reasons it is interesting to establish what features of speech are central in creating a voice impersonation that is convincing. Besides the entertainment aspect, security-demanding services protected by speaker verification systems may be vulnerable to mimics of a true client's voice. This poses a potential security problem and it is important to know how sensitive systems are and what can be done to improve their immunity to this type of fraud.

Spectral analysis has been used by Zetterholm (2003), who showed that, for instance, the professional impersonator adjusted his fundamental frequency and the formant frequencies of the vowels during impersonation to be closer to the target voice compared to that of his natural voice.

The ability of naive speakers and one professional impersonator to train their voices to a target speaker has been studied by Elenius (2001). In that work, the subjects could train their imitation by listening to repetitions of the target speaker and their own voice, and also by using the score of a speaker verification system as feedback. The false accept rate was significantly higher when the impersonators had trained their impersonation than before the training took place. This led to the conclusion that human impersonation is a threat to speaker verification. In the present report, we combine these two methods in order to study what features are used by the impersonator and how strong is their influence on the output score of the verification system.

A three-fold investigation has been conducted to investigate imitation success and to identify the core features for successful imitation. One, a speaker verification system was used: two, phonetic-acoustic measurements were made and three, a perception experiment was conducted. The first two issues have previously been addressed in (Blomberg, Elenius & Zetterholm, 2004). In addition to these, the current report includes the results of the listening experiments.

2. Speaker verification system

The speaker verification system used in this study is text-dependent and is similar to the one used by Melin, Koolwaaij, Lindberg and Bimbot (1998). A spoken utterance is segmented into separate words by a speech recogniser. Client and non-client (background) models are matched to the segmented speech. The background model has been trained by a number of non-client speakers. The logarithm of the ratio between the two matching scores, the log-likelihood ratio (LLR), is used as a verification score. A decision whether to accept or reject the claimed identity is taken, based on the verification score and a threshold.

The speech signal is sampled by 8 kHz, pre-emphasised and divided into 10 ms frames using a 25.6 ms Hamming window. Each frame is fed into an FFT-based, mel-warped, log-amplitude filterbank with 24 channels in the range from 300 to 3400 Hz. The filterbank spectrum is converted into 12 cepstrum coefficients and

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after abstract only review

one energy parameter. Their first and second time derivatives are included to a 39-component feature vector, which is input to the verification system.

One Hidden Markov Model (HMM) per word in the system vocabulary is used to model the pronunciation of each client. The number of states for each HMM is word-dependent and equals twice the number of phones in each word. A male and a female background model are trained using the database SpeechDat (Elenius and Lindberg, 1997). During verification, the male or female background model is chosen based on which seems most appropriate considering the speech signal.

3. Experiment

Experiments have been performed using a professional male Swedish impersonator speaking a four-digit sequence over a fixed-network, ISDN telephone connection. Recordings were made at three occasions: before having trained the impersonation using his natural voice, during the training session while adjusting his voice towards a target speaker, and after the completed training session during an attempt to maintain the impersonation without feedback. As feedback during training, three methods were used: audio playback of the target and the impersonation voices, the score of a speaker verification system, and a combination of these. Each training session was followed by a test session, which, in turn, was followed by a training session for the next feedback mode. The order between the feedback modes was kept constant, in the sequence described above. There was no constraint on the number of training attempts for any of the training modes. The recordings were analysed in order to measure voice differences before, during and after impersonation training. The speaker verification system was also used to score the success of the impersonations in all sessions.

In the experiment the four-digit sequence was kept fixed, 7, 6, 8, 9, in order to simplify the impersonation and the analysis.

4. Phonetic analyses

In order to understand how the impersonator succeeded in his imitations phonetic-acoustic measurements were made. For the acoustic analysis the Praat program (http://www.fon.hum.uva.nl/praat/) was used.

4.1. The impersonator

The male Swedish professional impersonator's dialect is a mix between a dialect from the western area of Sweden and a more neutral dialect. The impression is that he has an ordinary male pitch level and a sonorous voice quality. In all ten recordings with his natural voice, he pronounces the utterance as follows: [ʃʊː sɛks ɔta niːu] with short pauses between the digits. The articulation is distinct. The auditory impression of the intonation is that there is a slope with a higher pitch in the beginning of the utterance and the first digit is stressed.

4.2. The closest target voice

This male speaker's dialect is a central Swedish dialect, he has a rather low pitch level and sometimes a creaky voice quality, especially in the middle part of the utterance used in this study. He pronounces the four-digit sequence as follows: [ʃʊː sɛks ɔta niːɛ] without pauses, with a rather monotonous intonation and a slightly stressed last digit.

4.2.1. *The imitations*

The impersonator lowers his pitch level, uses a creaky voice quality in some parts of the imitations and changes his intonation pattern in order to get close to this target speaker. In some of the recordings he also changes his pronunciation of the last digit. However, according to the score, the verification system seems not to be very sensitive to this variation.

4.2.2. *The average F0*

Mean F0 was calculated based on measurements every 10 ms. The acoustic analysis of mean F0 confirms the auditory impression of a higher mean F0 in the recordings with the natural voice of the impersonator compared to this target speaker. See Table 1.

Table 1: Mean F0, std.dev. and score values for the impersonator's natural voice and the closest target speaker

Recording	Mean F0 (Hz)	Std.dev. (Hz)	Mean score
Natural voice, impersonator	125.8	35.3	-4.96
Target voice	119.0	9.1	-
Audio training	124.0	9.3	-1.97
Audio evaluation	113.9	6.0	0.18
Score training	113.9	6.6	-1.21
Score evaluation	119.9	5.9	-0.75
Audio+score training	119.9	5.9	-0.87
Audio+score evaluation	116.4	5.7	0.82

4.3. The median target voice

This male target speaker has a dialect from Stockholm, a low pitch level and a slightly nasal voice quality. He pronounces the four-digit sequence as follows: [ʃʊː sɛks ɔta niːu] without pauses and the first digit is slightly stressed. The articulation is not indistinct, but not as distinct as the impersonator.

4.3.1. *The imitations*

In the imitations of the median target speaker the impersonator lowers his own natural pitch level and changes his intonation. He also changes his own clear and distinct pronunciation towards the characteristics of this speaker.

4.3.2. The average F0

In this part of the experiment the impersonator has a lower mean F0 when speaking with his natural voice compared to the first part of the test. The acoustic analysis confirms the auditory impression of this speaker's low mean F0 and that the impersonator mean F0 is lowered in the imitations of this target speaker, see Table 2.

Table 2: Mean F0, std.dev. and score values for the impersonator's natural voice and the median target speaker

Recording	Mean F0 (Hz)	Std.dev. (Hz)	Mean score
Natural voice, impersonator	114.4	31.1	-6.96
Target voice	103.5	10.2	-
Audio training	104.2	8.6	-3.65
Audio evaluation	108.4	9.5	-3.26
Score training	106.8	11.0	-3.05
Score evaluation	111.6	12.1	-2.32
Audio+score training	102.7	13.9	-1.52
Audio+score evaluation	111.9	8.3	-1.81

There does not seem to be a strong relation between mean F0 and the score in any of the imitations of these two target speakers.

4.4. Vowel formants

A correlation analysis between the change in vowel formant frequencies and the score was conducted. The formants F1 through F4 were automatically tracked in the vowel segments using the Praat program and were manually corrected where necessary. Average frequencies were computed for each vowel. For relating the formant deviations with the verification system score, the frequency values were converted to mel scale. The reason for this is that the verification system uses this representation and comparisons will be more correct if performed in the same frequency scale.
The vowel distribution in the F1-F2 plane is plotted in Figure 1 for each target speaker, the impersonator's natural voice, and his evaluation recordings after the audio-score training. It is obvious that he adjusts his vowel positions for better, although not exact, correspondence with the target speakers.

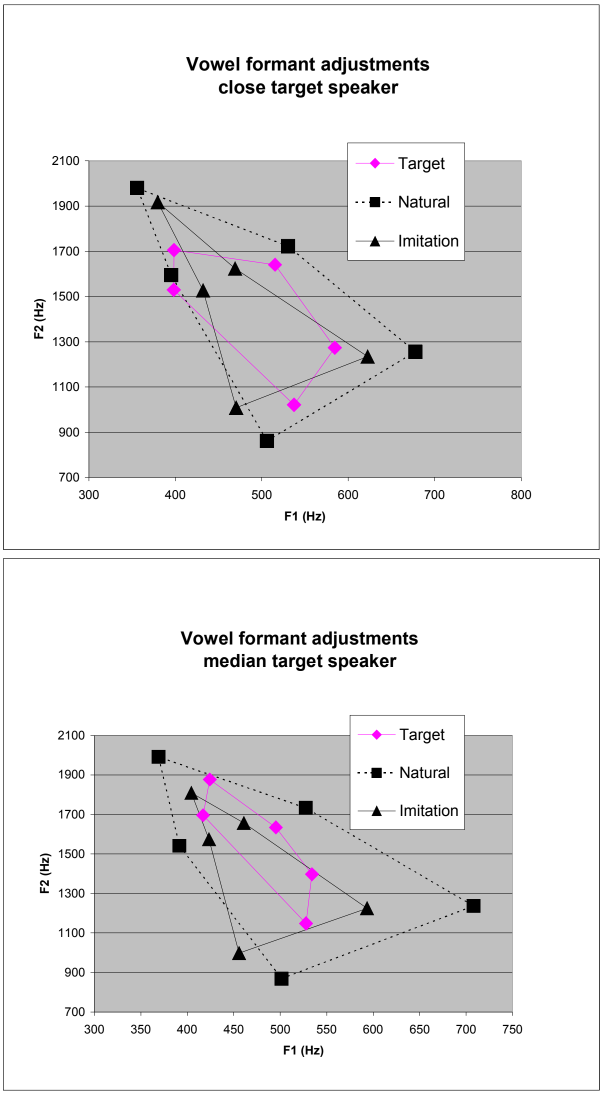

Figure 1: Vowel formant distribution for the close (top) and the median (bottom) target speakers and the impersonator's natural and mimic utterances.

Figure 2 shows the correlation between the formant deviation from the target speaker and the verification score of each utterance. All target speaker specific utterances (natural, training, and evaluation utterances) by the impersonator were used for this purpose. The pattern is similar for both target speakers. F2 has, as expected, a strong and negative correlation. F1 and F3 are less correlated. Preliminary analysis of F4 deviation have indicated a positive correlation with system score (Blomberg, Elenius & Zetterholm, 2004). Reliable conclusions for F4 require, though, higher bandwidth recordings than the 4 kHz used in this study.

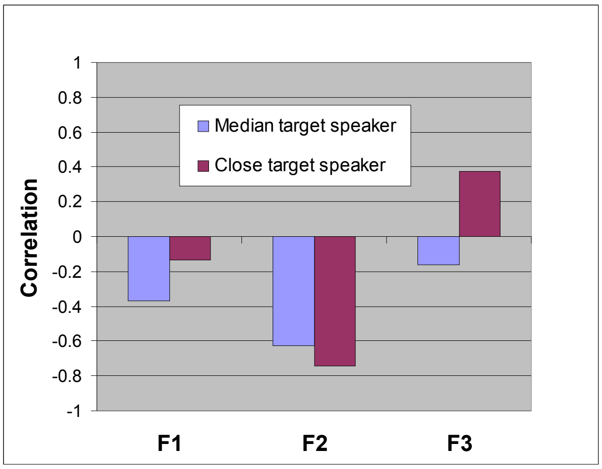

Figure 2: Correlation between the vowel formant magnitude deviation and the verification score.

Figure 3 shows a scatter diagram for the F2 deviation against verification score for the median target speaker.

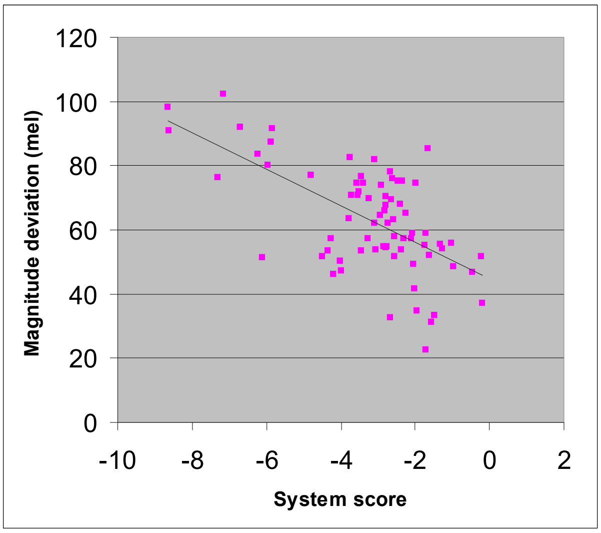

Figure 3: Scatter plot and regression line of the second formant magnitude deviation from the median target speaker against verification score.

5. Perception test

In order to ascertain whether human listener rank imitated voices in a similar manner to the speaker verification system, a perception test based on the recordings of the closest target speaker was designed.

5.1. The voices

One target utterance and 62 imitations were used. All speech segments were of the same four-digit sequence.

5.2. Design

An XAB test design was implemented in PsyScope (http://psyscope.psy.cmu.edu/). X was always the target voice and A and B were imitation utterances. 62 individual combinations of A and B were presented to each listener.

5.3. Listeners

22 listeners (12 male and 10 female, mean age 31) with no reported hearing problem undertook the perception test. All of the listeners were born in Sweden and are native speakers of Swedish yet with a range of different dialect backgrounds.

5.4. Procedure

The participants sat in front of a computer and listened to the stimuli through ear-phones. They were asked about their age, gender, dialect and whether they had any known hearing problem. They were instructed to respond A or B, depending on which of A and B was most similar to X. Prior to starting the experiment, a training phrase of six training pairs was undertaken. Then the participants were asked if they had any questions prior to starting the experiment.

5.5. Results

Figure 4 shows the agreement between the listeners' responses and that of the system, as a function of the magnitude difference between the system scores of the two utterances in each stimuli pair. The two histograms represent the number of agreeing and disagreeing judgments, respectively. For low and medium differences, the two histograms are essentially identical, indicating that the human and automatic decisions are independent in this interval. At higher system score difference, there is a tendency towards higher agreement between the listeners and the system. Still, linear regression only estimates 62% agreement for the stimuli pair with the highest system score difference.

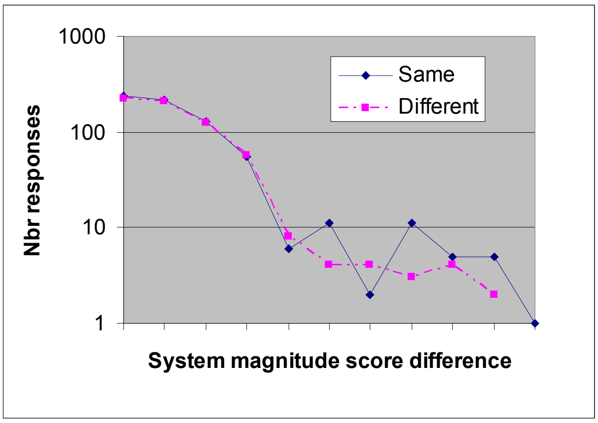

Figure 4: Number of agreements/disagreements between listener and system responses as a function of the score magnitude difference between the two imitation utterances.

By chance A and B happens to be the same sound file at a few occasions. Only 54% of all the answers in these cases are A, which means that there is no preference for the sound A even though this is presented before B and closest to the X sound.

5.5.1. Comments from the listeners

Most listeners comment that it sometimes was hard to hear the differences between voice A and B and to make a decision about which was most like X. When asked if they had any kind of strategy for their decision they tell that it often changed during the test. All listeners mentioned the different pronunciation of the last vowel between the X sound and some of the other utterances. In addition to that the pitch level, the prosody such as rhythm, pauses and the intonation seems important to the listeners.

6. Discussion

The results show that the impersonator really changes his natural voice and speech behaviour towards the two target voices. There are audible differences between the recordings, not only between the impersonator's natural voice and the two target voices, but also between the different voice imitations. Concerning the score of the recordings it is obvious that the impersonator is successful in his imitations, especially in the imitations of the first target speaker.

In the analysis of the vowel formants it is obvious that the impersonator adjusts his vowel positions to get closer to the target speakers. There is a particularly strong and negative correlation with score for the second formant, which indicates its high importance for a successful impersonation.

The results of the perception experiment show that the listeners agree with the system in their selection of the best of the two presented imitations around 60% of the time, when there is a large system score difference between the presented imitations. The agreement level drops rapidly as the system score difference decreases. Whether this indicates a system that is more sensitive than human speech perception or a human perception that is able to focus on specific elements of a recording to make better evaluations than this system is something that is currently unresolved and demands further consideration and investigation.

7. Conclusions

This comparison between human perception and a speaker verification system score of a voice imitation shows little agreement between listeners and the system. Imitations are evaluated differently by the system investigated in this paper and human listeners. The prosodic features, which seemed important to human listeners are not explicitly used by the system. The importance, if any, of this difference for the development of more secure systems warrants further investigation. The perception test placed large demands upon the listeners and it is possible that the listeners would have been more able to verify the correct voice in a standard verification test.

8. Acknowledgements

The research is funded partly by the Bank of Swedish Tercentenary Foundation through their funding of the project 'Imitated voices: A research project with applications for security and the law' and partly by the Vinnova national competence center Centre for Speech Technology (CTT), KTH, Sweden.

A special thanks to Joost van de Weijer for invaluable help with the design of the perception test. And thanks to all the listeners in Sweden.

9. References

Blomberg, M., Elenius, D., Zetterholm, E. (2004). Speaker verification scores and acoustic analysis of a professional impersonator. *Proc. Fonetik 2004: 84-87*, Dept. of Linguistics, Stockholm University, Sweden.

Elenius, D. (2001). *Härmning – ett hot mot talarverifieringssystem?* (in Swedish). Master thesis, TMH, KTH, Stockholm.

Elenius, K. Lindberg, J. (1997). SpeechDat Speech Databases for Creation of Voice Driven Teleservices. *Phonum 4, Phonetics Umeå*, May 1997:61-64.

Melin H., Koolwaaij J.W., Lindberg J., Bimbot F. (1998). A Comparative Evaluation of Variance Flooring Techniques in HMM-based Speaker Verification. *Proc. of ICSLP '98*: 1903-996.

Zetterholm, E. (2003). *Voice Imitation. A Phonetic Study of Perceptual Illusions and Acoustic Success.* Doctoral dissertation. Travaux de l'institut de Linguistique de Lund 44, Lund University, Sweden.