Jusczyk, Peter, Paul Luce & Jan Charles-Luce. 1994. 'Infants' sensitivity to phonotactic patterns in the native language'. *Journal of Memory and Language* 33, 630-45.
McQueen, James. 1998. 'Segmentation of continuous speech using phonotactics'. *Journal of Memory and Language* 39, 21-46.
McQueen, James, Dennis Norris & Anne Cutler. 1994. 'Competition in spoken word recognition: spotting words in other words'. *Journal of Experimental Psychology: Learning, Memory and Cognition* 20, 621-38.
Newsome, Mary & Peter Jusczyk. 1995. 'Do infants use stress as a cue in segmenting fluent speech?' In Dawn MacLaughlin & Susan McEwen (eds.), *Proceedings of the 19th Boston University Conference on language development.* Boston, MA: Cascadilla Press.
Smith, Neil. 1973. *The acquisition of phonology.* Cambridge University Press.
Weijer, Joost van de. 2003. 'Consonant variation within words'. In Dawn Archer, Paul Rayson, Andrew Wilson & Tony McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 Conference* (University Centre for Computer Corpus Research on Language Technical Papers, vol. 16), 184-90. Lancaster University, United Kingdom.

Joost van de Weijer <vdweijer@ling.lu.se>

Lund University, Dept. of Linguistics
Working Papers 51 (2005), 239–255

A case study of impersonation from a security systems point of view

Elisabeth Zetterholm, Daniel Elenius and Mats Blomberg

1 Introduction

Impersonation of a person, especially by means of voice, is sometimes used to amuse a human audience. Imitations often sound quite convincing. For several reasons it would be interesting to establish what aspects are important in performing a successful impersonation act. There are several components besides the acoustic signal that contribute to the subjective impression, such as the mood, the expectation, etc., and, if the impersonator can be seen, also visual similarity and gestural patterns. It may be difficult to determine the contribution of each individual factor.

It would also be interesting to use a more objective measure of performance than human impression. Objective measures of many of these aspects may be quite difficult to extract due to their complexity and lack of standardised analysis techniques. The acoustic signal, however, is suitable for objective evaluation, since there exist established techniques for quantitative phonetic analysis of speech and algorithms for determining the degree of acoustic similarity between utterances spoken by different persons.

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. In the present report, we complement these measurements with a computer-based speaker verification system. This type of system is normally used to judge, by the acoustic properties of a spoken utterance, whether a person has the identity (s)he has claimed or not. Our idea behind using this 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 analysis of the mimic speech.

2 Speaker verification

There are several occasions when a person is required to identify him/herself and to verify that (s)he has the claimed identity. In telephone and Internet banking, verification might be based on a combination of possessing a magnetic card and remembering a pin-code. At the entrance of a house, a four-digit code is sometimes used to unlock the gate. And a person might be requested to show his/her passport at the customs.

As codes may be hard to remember and artefacts might get lost, other techniques of verification have been sought for. One such technology is speaker verification. A speaker verification system accepts or rejects a claimed identity based on an analysis of how a phrase is spoken rather than its linguistic content. Therefore, the phrase does not need to be kept a secret. Thereby, the problem of forgetting a password may be avoided.

Speech is an act of performance. No human may perform identical pieces of speech twice; there are always small differences between two repetitions of an utterance. A decision must hence be made of how strict match is to be required between an access utterance and the reference model for the claimed identity. Thereby a trade-off between the risks of rejecting the true client and of accepting an impostor has to be made.

Before being able to use a speaker verification system, a new client has to be enrolled by speaking a number of utterances. These recordings are used to create a statistical model of the speaker. After the client has been enrolled, a person may claim having this identity and speak an utterance in order to access the protected service. The system then analyses the speech signal and makes a decision whether it accepts or rejects the claimed identity.

Fraud against security systems may be a problem. Impersonation is generally considered as a potential threat against a speaker verification system. Therefore, investigations of impersonation have been made. Both human impersonation and impersonation using technical aids, such as tape-recorders or speech synthesis systems trained on natural speech, have to be considered. A counter-plan against the tape-recorder may for instance be that the system randomly chooses what utterance to be spoken by the claimed client. This will make it more difficult to play-back a pre-recorded utterance from the true-identity speaker.

A method to counter a technical voice transformation technique was reported by Genoud & Chollet 1998. Lindberg & Blomberg 1999 have studied the vulnerability of a system against concatenative synthesis. The ability of naive speakers and one professional impersonator to train their voices to a randomly chosen target speaker has been studied by D. Elenius 2001. In this 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.

2.1 Verification method

Furui 1996 has given an overview of different methods for speaker verification. The most successful algorithms are based on the use of two statistical models. One model represents the claimed speaker (target speaker) and the other one is a non-speaker model (background model) representing any other speaker. Both models are matched against the access utterance and the decision is, in principle, a comparison between the likelihood that the true client or a background speaker would pronounce the utterance in this way. This comparison results in a score of the utterance. If the likelihood of the utterance given the client model is sufficiently much higher than that of the background model, a high score is given and the claimed identity is accepted.

An issue in speaker verification is how to set the threshold of accepting the claimed identity. If the threshold is high, few impostors are accepted. On the other hand, many true clients might not be accepted either. For a low threshold the opposite is true. In the process of choosing the threshold, the likelihood values of accepting impostors and rejecting true clients plays a central part. The costs of false rejection and false acceptance may also be specific for a particular application and the threshold can be adjusted to minimise the overall cost of the two types of errors.

The performance of a speaker verification system is normally determined by experiments on a speech database. The decision threshold is also set using separate speech data. In these corpora, speech is needed for clients as well as for impostors. The client speech data are simple collected by performing recordings of each speaker in his/her own voice. In contrast, databases with dedicated impostors, i.e. where the speakers are deliberately trying to enter the system using another client's identity, are quite rare. Therefore, performance of a speaker verification system is often measured using casual impostors. A casual impostor is a person who speaks with his/her natural voice but the utterance is used to access the account of another client in the system. In this case, an ordinary speech database may be used also for the impostors.

One criterion of choosing the threshold is to achieve an equal error rate (EER) of accepting impostors and rejecting clients. This criterion is often used in research. In a high security application, however, the resulting false accept rate might be too high. In an application that emphasises user acceptance rather than high security, the EER threshold may be too high. In these cases, the threshold may be chosen in order to minimise the total error cost on the particular database of speakers.

The Hidden Markov Model (HMM) approach is the dominating technique in speaker verification. This kind of model represents a signal source, i.e. a speaker, which, at each time point, may be in one of a set of states. In each of these states the source produces a statistically defined output. The output is often represented as spectral features of the speech signal. As the model goes through a sequence of states, a complex signal is produced, i.e. an utterance. If distinctly separable signals are generated in all states, it is possible to deduce the state sequence from the observed signal. However, in the general case this is not possible, hence the term *hidden*. Instead, one has to resort to calculate the probability of, for instance, the most probable state sequence. It is also possible to calculate the probability of the model having generated a particular observation. This measure is used to compute the verification score as was described at the beginning of section 2.1.

One representation of speech is as a sound wave in the time domain. The signal may also be characterised in the spectral domain. In this domain, frequency analysis of the signal may be performed. A third domain, the cepstrum, is reached by retransforming the logarithmic power spectrum into the time domain. The transformation and retransformation may be done by the fast Fourier transform, and its inverse. One motivation for using the cepstrum in speaker verification and speech recognition, is that the coarse spectrum shape can be represented by a few parameters (Gish & Schmidt 1994).

2.2 Difficulties in speaker verification

As was mentioned above, speech is an act of performance. Two utterances are never identical. Robustness against several sources of variability in the client's voice is needed. The overall speech level might differ between utterances. Another property is the fundamental frequency (F0), which might change due to the mental or physical state of the speaker. If the system is used in a telephone application, the microphone and transmission line may change from one

call to another. As these distort the speech in different manners, the speaker may sound different depending on which telephone is used to make a call.

Some distortion originating from the room acoustics, hand-set and transmission line may be reduced by time normalising the spectral parameters. This is achieved by subtracting the utterance average of each parameter from the parameter value in each time frame. Unfortunately, some characteristics of the voice are also lost during this procedure. Still, normalising the cepstrum was advantageous in an investigation by Bernasconi 1990. In this case, it was observed that the benefits of normalisation outnumbered the drawbacks.

Noll 1967 has shown that the vocal tract affects the low part of the cepstrum while the fundamental frequency generates a peak in the higher part of the cepstrum. Therefore, some robustness against the variability in F0 may be obtained by truncating the cepstrum below the F0 peak.

2.3 Technical specification of the system

The speaker verification system used in this study is similar to the one used by Melin et al. 1998. It consists of three parts: a speech recogniser, a speaker verification unit and a decision-making unit. Speech is segmented into digit-words by the speech recogniser. Speaker and word-dependent client and background models are matched to the segmented speech to produce a verification score. A decision whether to accept or reject the claimed identity is taken by the decision-making unit, based on the verification score and a threshold. In a practical application, knowledge of the requested identity of the utterance might also be used to check that the correct phrase was spoken.

The speech signal is pre-emphasised and divided into frames. A frame is created every 10 ms 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 3,400 Hz. A cosine-transform is applied to the filterbank signal to create 12 cepstrum parameters. These parameters and an energy component are used together with their first and second time derivatives to form a 39-component feature vector, which is used by the verification system. The energy component is formed from the 0th cepstrum component and represents the raw log-energy in each frame. This component is scaled to have the same maximum for each utterance.

One HMM per digit is used to model how the client speaks the digits from zero to nine. The number of states for each HMM is digit-dependent and is equal to twice the number of phones in each digit. The client models were trained on 25 five-digit utterances stored in the *Gandalf* database (Melin

1996). A male and a female background model were trained using another database, *SpeechDat* (K. Elenius & Lindberg 1997, K. Elenius 1999). 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 impersonator speaking a four-digit sequence over a fixed-network telephone connection. Recordings were made at three occasions: before having trained the impersonation, using the impersonator's 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. 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 the experiment, a fixed four-digit sequence, 7 6 8 9, was used to simplify the impersonation and the analyses. Two target voices were used in the experiment. These where selected based on scoring 10 of the impersonator's natural utterances against the reference models of 22 target speakers. The average score against each target was calculated, and the two target speakers with the highest and the median score, respectively, were chosen. Both speakers were male.

The experiment was carried out using the same procedure and graphical user interface as in D. Elenius 2001. The subject indicated the start of each utterance by clicking on a screen button. An end-point detector decided when the utterance was finished, after which segmentation into digits was performed by a speech recogniser. If the recogniser did not recognise the prompted identity, a warning was shown and the subject had to repeat the recording.

The recordings of the subject's natural voice and the recordings after each training session were made with similar graphical interfaces. In these, no feedback of the similarity to the target speaker was given. As ten repetitions of the utterance were requested, a progress meter was used to show the current number of correctly recognised utterances. Utterances that were misrecognised, i.e., recognised as something else than 7 6 8 9, were discarded and a note about what was recognised was shown.

Table 1. Verification scores for the impersonator's mimic of the two target voices and different training modes.

Session	Closest target speaker			Median target speaker		
	N	mean	st. dev	N	mean	st. dev.
Natural voice	10	–4.96	0.62	10	–6.96	1.05
Audio training	6	–1.97	1.93	16	–3.65	1.25
Audio evaluation	10	0.18	0.50	10	–3.26	0.44
Score training	12	–1.21	0.76	22	–3.05	0.72
Score evaluation	10	–0.75	0.71	10	–2.32	0.88
Audio+score training	13	–0.87	0.58	13	–1.52	0.99
Audio+score evaluation	10	–0.82	0.90	10	–1.81	0.83

Three training modes were used. Each training session was followed by a test session. There was no constraint on the number of training attempts for any of the training modes.

In the first training mode, the subject was able to listen to the target voice, make a recording of his own attempt at imitating the target, and then listen to this attempt. Each of these choices was represented by a button on the screen, and the selection was done by clicking with the mouse.

In the second training mode, the similarity to the target voice, i.e. the score from the speaker verification system, was shown as the vertical position of a point in a graph. The subject was able to click on the points on the graph to play the corresponding recording. Thereby, the subject might deduce that a particular change in the voice increased or decreased the score.

In the third training mode, it was possible both to listen to the target voice and to get the graphical and audio feedback mentioned in the previous paragraph. This mode was intended to enable the subject to fine-tune his mimic based on both listening and verification system feedback.

4 Verification results

The verification scores of the two target voices and the three training modes are shown in Table 1.

It is obvious that the impersonator is successful in imitation, especially the closest target voice. The highest score is found in the part of the test with feedback both with playback of the target voice and the impersonation. All imitations in this part have very high scores. There are only small differences in score between all imitations of this target speaker, in general.

When imitating the median target speaker, the impersonator got the highest score when using both audio and verification score as feedback during training. Generally, there are greater differences between the imitations of the second target speaker, according to the score by the speaker verification system, compared to the imitations of the first target speaker.

A correlation analysis between the change in the measured characteristics and the score was conducted. The measured characteristics were: average F0, the F0 contour for both target voices, and formant frequencies, F1-F4, of the different vowels. In this paper the correlation between F0 and the score will be discussed. The analysis between the change in vowel formant frequencies and the score is presented in Blomberg, Elenius & Zetterholm 2004.

5 Phonetic analysis

In order to understand how the imitator succeeded in his imitation it is necessary to find out which phonetic features in his voice that resulted in the system accepting his voice.

5.1 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. Ten recordings with the four-digit sequence were made with his natural voice at the beginning of each test with different target voices. In all recordings he pronounces the utterance as follows: [ʃʉ: sɛks ɔta ni:u] with a short pause between every digit in the recordings for the first test with the closest, highest scored, target voice, but not before the second test with the median scored target voice. The speech tempo is lower in the first ten recordings. The articulation is distinct. The auditory impression of the intonation is that there is a slope with a higher pitch at the beginning of the utterance and the first digit is stressed. The same intonation pattern occurs in the recordings before the test with the first target speaker, as well as before the test with the second target speaker. This indicates that the influence from the first target speaker on the impersonator's own natural speech before the second speaker is of minor importance concerning the intonation pattern.

5.2 The closest target voice

The auditory impression of this speaker is that he speaks with 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 as the target utterance in this study. He pronounces the four-digit sequence as follows: [ʃʉ: sɛks ɔta ni:ə] without pauses, with a rather monotonous intonation and a slightly stressed last digit.

5.2.1 The imitations. When imitating this target voice, the impersonator lowers his pitch level and in some of the recordings he changes his pronunciation of the last digit. Sometimes he uses a creaky voice quality and he changes his intonation to get close to the target speaker.

Concerning the pronunciation of the last digit, the impersonator changes his own natural pronunciation in the three first parts of the test, but in the fourth part there is a variation between [ni:ə] and [ni:u] with a transition to his own pronunciation. In the two last parts of the test he uses his own natural pronunciation [ni:u] in all recordings. However, judging from the score, the verification system seems not to be that sensitive to this variation.

5.2.2 The average F0. For the acoustic analysis the Praat program (http://www.fon.hum.uva.nl/praat/) was used. Mean F0 was calculated based on measurements every 10 ms.

The acoustic analysis confirms the auditory impression of a higher F0 in the recordings with the natural voice of the impersonator compared to the selected target voice. The ten recordings with the impersonator shows a mean F0 of 126 Hz, and the target speaker has a mean F0 of 119 Hz. The best-scored recording with the impersonator's natural voice has a mean F0 of 121 Hz, which is rather close to the target speaker. The imitation with the highest score has a mean F0 of 114 Hz (Table 2). There does not seem to be a strong relation between F0 and the score.

Table 2. Mean F0, standard deviation and score values for the impersonator's natural voice and the highest scored target voice.

	mean F0 (Hz)	st. 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	6.8	–0.75
Audio+score training	119.9	5.9	–0.87
Audio+score evaluation	116.4	5.7	–0.82

However, the differences in standard deviation between the recordings are considerably more interesting. The impersonator's natural voice has a standard deviation of 35 Hz, as an average of all ten recordings as well as in the best scored recording. In the target speaker recording, the standard deviation is only 9 Hz. This difference probably accounts for the auditory impression of the differences in intonation between the two voices. The imitation with the highest score has a standard deviation of 6 Hz, which clearly shows the difference to the natural speech of the impersonator.

The imitations in different parts of the test show a variety of mean F0, but it is obvious that the impersonator tries to lower his own fundamental frequency and the frequency range in all imitations and that he is very close to the target speaker.

5.2.3 The F0 contour. The acoustic analysis of the F0 contour confirms the auditory impression of a downward slope in the recordings with the impersonator's natural voice. The primary stress is on the first digit in this utterance (Figure 1).

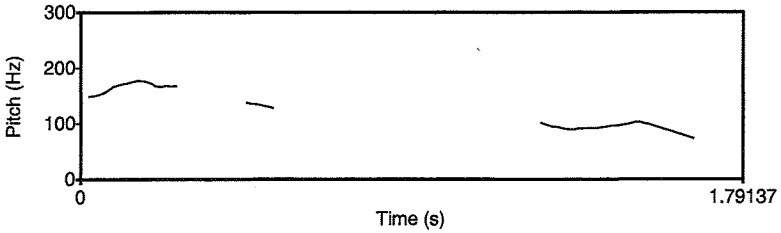

Figure 1. The F0 contour of the impersonator's own natural voice.

In the recording with the target voice there is less variation in the intonation contour, which confirms the auditory impression. The last syllable is slightly stressed with an increased F0 (Figure 2).

When comparing the F0 contour of the target voice with one of the best scored imitations it is obvious that the impersonator really changes his natural F0 contour and is very close to the intonation of the target speaker. He has lowered his F0 and the intonation is rather monotonous with an increased F0 on the last syllable (Figure 3).

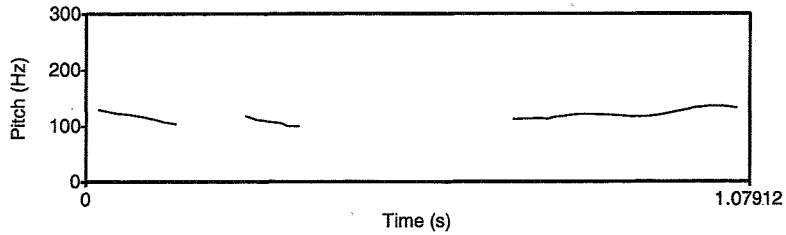

Figure 2. The F0 contour of the highest scored target voice.

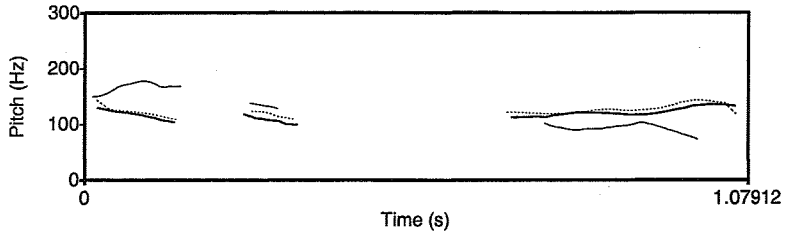

Figure 3. The F0 contour of the impersonator's own voice, the highest scored target voice (bold) and the best scored imitation (dotted line).

5.3 The median target voice

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

5.3.1 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 in order to sound more like this target voice.

5.3.2 The average F0. In the acoustic analysis it is obvious that the impersonator has a lower mean F0 speaking with his own natural voice in these recordings compared to the recordings in the test with the first, highest scored target voice. The average F0 of these ten recordings is 114 Hz. The best-scored recording with the impersonator's own voice has an average F0 of

Table 3. Mean F0, standard deviation and score values for the impersonator's natural voice and the median scored target voice.

	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

120 Hz. The target speaker has an average F0 of 104 Hz. The imitation with the highest score has a mean F0 of 103 Hz (Table 3).

The standard deviation in F0 for the impersonator's natural voice has an average of 31 Hz, and is 33 Hz in the highest-scoring recording. There is a difference between the impersonator and the target voice in this test, as in the first test. The target speaker has a standard deviation of 10 Hz. The measurements of all imitations show a low frequency range close to the target speaker. The highest-scoring imitation has a standard deviation of 14 Hz.

There are differences in mean F0 between all the recordings in this test, but it is clear that the impersonator lowers his natural mean F0 and the F0 range in the imitations and that he is closer to the target voice than to his own natural voice.

5.3.3 The F0 contour. The F0 contour of the impersonator's own voice shows a similar pattern as in the first test, with a down-sloping intonation contour (Figure 4).

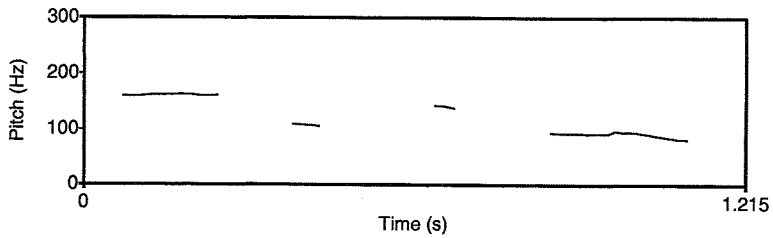

Figure 4. The F0 contour of the impersonator's own natural voice.

The impression of a rather monotonous intonation is confirmed in the acoustic analysis of the second target speaker. In this recording, the first digit is slightly stressed. A somewhat increased F0 is seen at the last syllable (Figure 5).

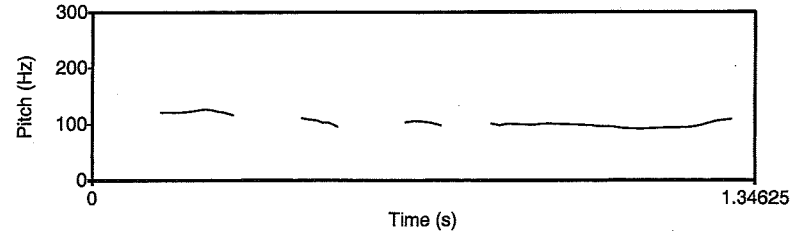

Figure 5. The F0 contour of the second median scored target voice.

Figure 6 shows the comparison between the recordings of the target voice (bold), the best scored imitation (dotted line) and the impersonator's own natural voice. The imitation is rather close to the intonation contour of the target speaker and it is obvious that the impersonator really changes his own F0 contour, at least in the beginning of the utterance. All F0 contours are rather close at the end of the four-digit sequence, except from the last syllable, which shows an increased F0 in both the target voice and the imitation. When speaking with his natural voice, the impersonator has a decreased F0 at the last syllable.

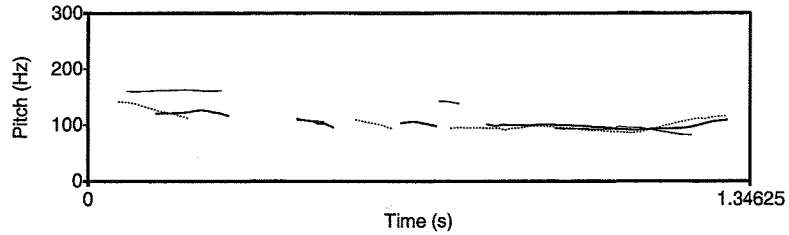

Figure 6. The F0 contour of the impersonator's own voice, the median scored target voice (bold) and the best scored imitation (dotted line).

6 Discussion

The results show that this impersonator really changes his natural voice and is able to imitate the two target voices with success. 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 of the target voices. The impersonator changes his pitch level, the intonation and his pronunciation. In the acoustic analyses it is obvious that he lowers his natural F0 in all recordings. There is a great difference between the F0 range in the recordings with the impersonator's own voice and both target voices. The measurements of F0 indicate that he is aware of the differences and changes his own speech behaviour in order to get close to the target voices. The acoustic analyses of the F0 contour confirm the differences and show that the impersonator changes his natural intonation in both tests. The down sloping F0 contour in the recordings with his natural voice is changed into a more monotonous contour and F0 is increased on the last syllable, which is an opposite pattern compared to his natural speech. See Figures 3 and 6.

Both the impersonator and the two target speakers have normal male voices without any specific characteristic features. The first target speaker has a creaky voice quality in the middle of his recording and the second target speaker has a slightly nasal voice quality. The impersonator does not imitate these voice qualities to a great extent in any of his recordings and there is no correlation to the score values. That may indicate that the verification system is not sensitive to the different voice qualities in these recordings. However, since the impersonator has a sonorous voice quality, the system did not select a voice with a distinguishing voice quality. Nevertheless, it is interesting to notice that the creaky voice quality in the recordings with the closest target voice is very obvious, both in the auditory and the acoustic analyses.

Concerning the score of the recordings is it obvious that the impersonator is successful in his imitations, especially in the imitations of the first target speaker. In the second test the score values are not that high, but still it is clear that the imitations are very close to the target speaker in most of the recordings.

The different feedback modes during training should not be compared in this experiment, since their sequence was fixed and a learning effect is likely to have appeared. To evaluate their importance, a higher number of impersonator and target speakers need to be included with balanced mode ordering. Nevertheless, in this experiment, there were several observations made that would be interesting to test on a larger material.

When comparing the score and the mean F0 it is hard to find a clear relation between a mean F0 very close to the target speaker and a high score. Some of the imitations with a low score show a mean F0 closer to the target speaker than some of the imitations with a higher score. This corresponds to expectations, since the verification system does not include explicit measurements of fundamental frequency. The same pattern occurs when comparing the standard deviation between the utterances and the score values. The results of the F0 measurements show that the impersonator gets close to the F0 of the target speaker, but this is not a very important feature in the system's identification procedure.

A clear difference between the standard deviation of F0 in the recordings with the impersonator and the target speakers is shown in the phonetic analyses, both the auditory impression and the acoustic measurements. In all voice imitations it is obvious that the impersonator changes his standard deviation to a great extent. The impression of monotonous speech in the recordings with the target speakers, compared to the impersonator's speech, is confirmed in the acoustic analyses.

The recordings of the four-digit sequence show a great difference in the intonation contour between the impersonator and the target speakers. The impersonator has a clear down-sloping intonation contour and both target speakers have more monotonous intonation with a slightly stressed last syllable and increased F0 at the end of the utterance. The median target speaker stresses the first digit as well. When comparing the recordings of the closest target speaker and the best scored imitation it is obvious that the imitation is very close to the intonation contour of the target speaker, see Figure 3. There is a clear difference between the recording with the impersonator's natural voice with down-sloping intonation and this imitation. In fact, all imitation recordings exhibit this intonation contour, regardless of their score values. The intonation may be an important feature in the voice identification task for human listeners, and it is likely that it could be of value for a speaker verification system as well.

In forensic phonetic research the formant frequencies of the vowels seem to be important for comparison between recordings and as speaker identity cues (Hollien 2002, Rose 2002). Rose mentions that formants are not explicitly used in automatic speaker recognition systems. The study of the relation between the vowel formant frequencies in the different recordings in this study and their relation to the verification score show a strong correlation, especially for

the second formant, which indicates its high importance for a successful impersonation (Blomberg, Elenius & Zetterholm 2004).

7 Concluding remarks

Both the score and the measured characteristics show that it is possible to change one's voice and speech behaviour and get close to another speaker. The results indicate that there is a possibility to trick both human listeners and a speaker verification system. Probably there are many different features which are important both for the voice identification task and the speaker verification system. To do a general ranking list of these features may not be possible. Since almost all imitations are very close to the target speakers and the impersonator changes more than one feature in his voice and speech it is hard to know the importance of a single feature alone. Further studies and comparisons between the recordings used in this study may give an insight into these important features.

Interesting follow-up work will be to relate listener's auditory impression of a successful voice imitation with the score of the speaker verification system.

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 Center for Speech Technology (CTT), KTH.

References

Bernasconi, C. 1990. 'On instantaneous and transitional spectral information for text-dependent speaker verification'. *Speech communication* 9:2, 129-39.

Blomberg, M., D. Elenius & E. Zetterholm. 2004. 'Speaker verification scores and acoustic analysis of a professional impersonator'. *Proceedings, Fonetik 2004,* 84-87. Dept of Linguistics, Stockholm University.

Elenius, D. 2001. *Härmning: ett hot mot talarverifieringssystem?*. Master thesis, Department of speech, music and hearing, KTH, Stockholm.

Elenius, K. 1999. 'Experiences from building two large telephone speech databases for Swedish'. *TMH-QPSR* 1-2/1999, 51-56.

Elenius, K. & J. Lindberg. 1997. '*SpeechDat* speech databases for creation of voice driven teleservices'. *Phonum* 4, 61-64.

Furui, S. 1996. 'An overview of speaker recognition technology'. In C.-H. Lee, F. K. Soong & K. K. Paliwal (eds), *Automatic speech and speaker detection*, 31-56. Boston: Kluwer Academic Publishers.

Genoud, D. & G. Chollet. 1998. 'Speech pre-processing against intentional imposture in speaker recognition'. *Proceedings of the International conference on spoken language processing 1998,* 3: 1211-14.

Gish, H. & M. Schmidt. 1994. 'Text-independent speaker identification'. *IEEE Signal processing magazine*, Oct. 1994, 18-32.

Hollien, H. 2002. *Forensic voice identification.* London: Academic Press.

Lindberg, J. & M. Blomberg. 1999. 'Vulnerability in speaker verification: a study of technical impostor techniques'. *Proceedings of Eurospeech'99,* 3: 1211-14.

Melin, H. 1996. '*Gandalf:* a Swedish telephone speaker verification database', *Proc. Fourth International conference on spoken language processing (ICSLP'96)*, Philadelphia, 1954-57.

Melin, H., J.W. Koolwaaij, J. Lindberg & F. Bimbot. 1998. 'A comparative evaluation of variance flooring techniques in HMM-based speaker verification'. *Proceedings of the 5th International conference on spoken language processing,* 1903-06.

Noll, A. M. 1967. 'Cepstrum pitch determination'. *The journal of the acoustical society of America:* 293-309.

Rose, P. 2002. *Forensic speaker identification.* London: Taylor & Francis.

Zetterholm, E. 2003. *Voice imitation: a phonetic study of perceptual illusions and acoustic success* (*Travaux de l'institut de linguistique de Lund* 44). Lund University.

Praat software package: http://www.fon.hum.uva.nl/praat

Elisabeth Zetterholm <Elisabeth.Zetterholm@ling.lu.se>
Daniel Elenius, Dept. of Speech Music and Hearing, KTH, Stockholm
<daniele@speech.kth.se>
Mats Blomberg, Dept. of Speech Music and Hearing, KTH, Stockholm
<mats@speech.kth.se>