Speaker verification scores and acoustic analysis of a professional impersonator

Mats Blomberg¹, Daniel Elenius¹, Elisabeth Zetterholm² *
¹Department of Speech, Music and Hearing, KTH, Stockholm
²Department of Philosophy and Linguistics, Umeå University, Umeå

Abstract
A professional impersonator has been studied when training his voice to mimic two target speakers. Training was performed by listening and/or using the output score from a speaker verification system as feedback. Tests after training showed a significant increase in verification score. An acoustic analysis shows how the impersonator adjusts his formant positions for the target speaker. Strong correlation with verification score is shown for the second formant.

Introduction
Impersonation of a person, and especially by means of voice, is sometimes used to amuse a human public. Imitations often sound quite convincing. For several reasons it would be interesting to establish what aspects are important in performing a successful impersonation act. Besides the entertainment aspect, security-demanding services protected by speaker verification systems may be vulnerable to mimics of a true client’s voice. This may be a security problem and it is important to know how sensitive the systems are and what can be done to improve their immunity to this type of fraud.

Analysis of a professional impersonator has been performed by Zetterholm (2003), who showed that, for instance, the impersonator adjusted his fundamental frequency and the formant frequencies of the vowels during impersonation to approach the target voice.

The ability of naive speakers and one professional impersonator to train their voices to a randomly chosen 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 acceptance 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 their influence is on the output score of the verification system.

Verification system

The speaker verification system used in this study is text-dependent and is similar to the one used by Melin, Koolwaaj, Lindberg and Bimbot (1998). The verification utterances are restricted to digit sequences. A spoken utterance is segmented into digits 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 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 digit is used to model the pronunciation of each client. The number of states for each HMM is digit-dependent and equals twice the number of phones in each digit. A male and a female background model were 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.

¹ Names in alphabetic order
Experiment
Experiments have been performed using the same male Swedish professional impersonator as in Zetterholm (2003). The utterance to be mimicked was a four-digit sequence spoken over a fixed-network 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: (1) audio playback of the target and the impersonation voices, (2) the score value of a speaker verification system, and (3) 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 above sequence. There was no constraint on the number of training attempts for any of the training modes. The experiment was carried out using the same procedure and graphical user interface as in (Elenius, 2001).

In all sessions, the speaker verification system was used to score the success of the impersonations. The recordings were also analysed acoustically in order to measure voice differences before, during and after impersonation training.

In the experiment, the four-digit sequence was kept fixed ("7 6 8 9") in order to simplify the impersonation and the analysis. Two target voices were used in the experiment. These where selected based on scoring 10 utterances of the impostor’s natural voice against a set of 22 speakers. Their models had previously been trained on 25 five-digit utterances. The average score against each speaker was calculated, and the two speakers with the highest and the median score, respectively, were chosen as targets for the experiment. Both speakers were male.

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 was successful in imitation, especially to the closest target voice. All imitations in this part have very high scores. When imitating the median target speaker, the impersonator increased his score gradually during the experiment. It is not possible to determine the best training mode due to the low number of target speakers and also since the fixed sequence of the training modes does not allow for normalisation of training effects between the sessions.

Table 1. Average verification scores for the impersonator’s mimic of the two target voices and different training modes.

<table>
<thead>
<tr>
<th>Session</th>
<th>Closest target speaker</th>
<th>Median target speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural voice</td>
<td>-4.96</td>
<td>-6.96</td>
</tr>
<tr>
<td>Audio training</td>
<td>-1.97</td>
<td>-3.65</td>
</tr>
<tr>
<td>Audio evaluation</td>
<td>0.18</td>
<td>-3.26</td>
</tr>
<tr>
<td>Score training</td>
<td>-1.21</td>
<td>-3.05</td>
</tr>
<tr>
<td>Score evaluation</td>
<td>-0.75</td>
<td>-2.32</td>
</tr>
<tr>
<td>Audio+score train.</td>
<td>-0.87</td>
<td>-1.52</td>
</tr>
<tr>
<td>Audio+score eval.</td>
<td>-0.82</td>
<td>-1.81</td>
</tr>
</tbody>
</table>

Phonetic analysis
The impersonator
The 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 recordings with his natural voice, he pronounces the utterance as follows: [SU: sEks Ota ni:u] with short pauses between the digits. The articulation is distinct.

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 target utterance in this study. He pronounces the four-digit sequence as follows: [SU: sEks Ota ni:E] without pauses, with a slightly stressed last digit.

The imitations
When imitating this target voice, the impersonator lowers his pitch level. Sometimes he uses a creaky voice quality and he changes his intonation to get close to the target speaker. In some of the recordings he changes his pronunciation of the last digit. However, according to the score, the verification system seems not to be very sensitive to this variation.
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 follows: [SU: sEks Ota ni:u] without pauses and the first digit is slightly stressed. The articulation is not indistinct, but not as distinct as that of the impersonator.

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.

Acoustic analysis

A detailed analysis of the fundamental frequency is given in Zetterholm et al (2004). Here, we complement that report with results of the formant analysis.

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. 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 those of the target speakers.

To have better understanding of the importance of the formant adjustments for the verification score, the correlation was measured between the average magnitude formant deviation in an utterance and its verification score value. All target speaker specific utterances by the impersonator were used for this purpose. Figure 2 shows the correlation for the individual formants. The pattern is quite similar for both target speakers. F2 has, as expected, the strongest correlation. F1 and F3 are less correlated. Surprisingly, the second most correlated formant is F4. Still more unexpected is the fact that the correlation is positive, i.e., the system similarity judgment increases with larger F4 distance between the impersonator and the target speaker. This holds for both speakers.

1 http://www.fon.hum.uva.nl/praat/
score is positive. One explanation might be that F4 is actively controlled to compensate for mismatch in other parameters, such as voice source spectrum or vocal tract length. Another possibility could be that F4 has little impact by itself but covaries with a feature having strong score correlation, e.g. F2. Finally, there is also a non-negligible risk of formant tracking errors for some utterances. This will be investigated by manual inspection. The correlation is close for the two target speakers, but more speakers are required to determine whether it is a general behaviour of this impersonator.

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


