# Same speaker – different voices
# A study of one impersonator and some of his different imitations

**Elisabeth Zetterholm**

Centre for Languages and Literature,
Linguistics/Language Technology
Lund University, Sweden
elisabeth.zetterholm@ling.lu.se

## Abstract

One impersonator and a number of his different voice imitations have been studied in order to gain some insights into the flexibility of the human voice and speech. A whole text unit and one selected word, which occur in all the recordings, have been analyzed and a comparison made between the different recordings. Both auditory and acoustic analyses have been attempted. The results indicate that this impersonator is able to adopt a range of articulatory-phonetic configurations in order to achieve the target speakers. This very fact raises questions concerning features that are hard to change in the voice and the possibility to find some kind of long-term signature for one speaker.

## 1. Introduction

Different studies of professional impersonators and their voice imitations show that it is possible to get close to another speaker's voice and speech behaviour, both in perceptual and acoustic analyses (Zetterholm, 2003). An impersonator is used to resembling other people and has the ability to pretend and make other people believe that they are another person. In order to imitate a certain target speaker, the impersonator has to select and copy many different features, laryngeal as well as supralaryngeal, to be successful and to convince the listeners about the target speaker. Focusing on important features and passages in the text, as well as exaggerating characteristic features, may be a conscious way of working with and improving the impersonation.

There are organic differences between speakers, which cannot be changed, making it hard to produce exact copies of another speaker's voice and speech. As a result, imitation is often a stereotyping process (Laver, 1994). The reader will probably think that a caricature is more entertaining, which is often the purpose when performing on stage. Nevertheless, it has been shown that high-quality voice imitation can mislead the listener (Schlichting & Sullivan, 1997). In regard to imitations of a familiar voice, it has also been shown that the listener may have expectations about characteristic features of the target speaker's voice and speech behaviour. Using words and phrases related to the target speaker make it easier for the audience to identify the imitated voice (Zetterholm et. al, 2002). A related observation was made in a recent comparative study of imitations produced by two professional impersonators and one amateur. The professionals were indeed more flexible in their imitations than the amateur, both concerning voice quality, mean F0 and articulation rate (Zetterholm, 2006). Despite these differences, the listeners agreed that they recognized the imitated voices.

The type of evidence outlined above does suggest that security-demanding services protected by speaker verification systems, for example, are likely to be vulnerable to mimics of a true client's voice. It will therefore be important to know how sensitive the systems are and what can be done to improve their immunity to this type of fraud. 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). 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 could indeed be a threat to speaker verification. One would expect the robustness of automatic systems to be also threatened by similar-sounding voices. Elliot's (2000) and Rose's (1999) related investigations of the words ''okay'' and ''hello'', respectively, have shown that similar-sounding voices can yield significant differences between different speakers' F (Formant)-patterns.

To disentangle the problems arising from similar-sounding voices, through impersonation or not, it will be critical to account for the quality of a speaker's voice, which depends on organic and learned behaviours (Laver, 1994). Between-speaker variation is partly due to organic differences and may vary from minor differences such as dentition to greater differences such as overall size and shape of the vocal organs. Learned differences start when children acquire the social and regional dialect of their environment. There are long-term organic characteristics as well as momentary changes, which influence our speech. Within-speaker differences occur during normal speaking conditions e.g. due to state of health, emotions, the speaker's attitude and the communicative intention, and indicate that the human voice is

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. 

Accepted after full paper review

very flexible. Despite these differences we can recognize familiar voices, even though we sometimes make mistakes.

To imitate another speaker's voice and speech behaviour, it is implicitly understood that an impersonator has to be aware of not only different markers of group identity such as regional or social dialect (Pittam, 1994, *inter alia*), but also personal markers in speech such as pronunciation or articulation (studied among others by Scherer, 1979). It is also presumed that the impersonator has to be able to change his own articulatory settings to get close to another speaker's voice and speech behaviour. This study is a first attempt at elucidating the voice flexibility exhibited by one impersonator by seeking evidence from his imitations, which would yield some insights into his articulatory-phonetic strategies. Such insights should ultimately be useful in further research about voice identification.

## 2. The present study

### 2.1. Material

One Swedish professional impersonator recorded a tape for the Swedish telephone company Telia, containing imitations of some well-known Swedish people. Some of them are politicians, others are well-known TV-hosts or newsreaders. Twelve (12) voice imitations have been selected for this study.

The texts have been created, by the impersonator, to suit typical vocabulary use and the features of each target speaker. The idea is that this should make it easier for the listener to recognize the voice. These recordings were designed to be humorous in the first place. All of the recordings contain one word (*mobilsvar*: the mobile phone answering machine), which is not a typical or even a frequently used word for any of the target speakers. This word was thus selected for comparison between the different voice imitations since it occurs in all recordings. There is also one recording of the selected word with the impersonator's own natural voice and speech, made for comparison with the imitations. The basic rhythmic structure of the word *mobilsvar* consists of three syllables, unstressed *mo*, primary stress on *bil* and secondary stress on *svar*.

#### *2.1.1. Dialectal differences*

The dialects of the original voices are different with regard to both segments, such as the r-segment, and the intonation pattern. None of the target speakers has the same dialect as the impersonator. He has a dialect from the eastern part of Sweden. The target speakers have dialects from the western, eastern or southern part of Sweden, some of them influenced by the dialect of Stockholm. One target speaker has a clear Stockholm dialect and two of the speakers have a clear dialect from the city of Gothenburg. There is considerable variation between Swedish regional dialects concerning both phonetics and phonology. E.g. there are two main types, [r] and [ʀ], of the phoneme /r/. The most common form is the alveolar trill [r] that is used in the majority of the Swedish dialects outside the South of Sweden. In southern Swedish dialects a uvular fricative [ʁ] or a uvular trill [ʀ] is used. Various diphthongizations of long monophthongs are one characteristic dialectal marker of the southern dialects, often with an initial onglide to the target vowel. The short vowels are often monophthongs (Bruce, 1970; Elert, 1991).

### 2.2. Method

An auditory and an acoustic analysis have been undertaken in order to see how, and to what extent, the impersonator changes his own voice and speech behaviour in these imitations. Comparisons were made between the different voice imitations and between the imitations and the natural voice of the impersonator. No comparison was conducted with recordings of the target speakers.

A perceptual evaluation of the various recordings was conducted by three Swedish well-trained phoneticians, who were familiar with the target voices. The analysis programme Praat (http://www.fon.hum.uva.nl/praat/) was used for the acoustic analyses.

In the auditory analysis, pitch, voice quality, dialect, articulation and individual phonetic habits and characteristic features were scrutinized. In this paper, however, only the general impression will be presented (for further details, see Zetterholm, 2003). A phonetic transcription of the word *mobilsvar* was made by the three phoneticians.

The acoustic analysis consisted of measurements of the mean fundamental frequency (F0), for the whole utterance as well as for the selected word, *mobilsvar*. Duration measurements were made for the whole text unit, the selected word and the stressed vowels /i:/ and /a:/. An estimate of articulation rate (in syllables per second) was calculated for the whole text. Formant frequencies of the stressed vowels /i:/ and /a:/ in *mobilsvar* were obtained as well. All measurements were made for the impersonator's own voice and his 12 imitations.

## 3. Results and comments

The results of the auditory and the acoustic analysis refer to the whole text unit as well as to the selected word, *mobilsvar*.

### 3.1. Auditory analysis

The general auditory impression is that this impersonator shows a great deal of intonational flexibility by emulating widely-varying pitch levels. He is rather successful in imitating the different voice qualities associated with the target speakers. Some speak with a nasal voice quality, others with a creaky or a tense voice quality, or with a combination of these. Two of the imitations render well the tense voice quality that is usually perceived in certain target speaker's deliveries of political speeches.

The impersonator captures the different Swedish dialects very well with respect to intonation patterns and different pronunciation of the sound segments, the r-segment, the s-segment as well as vowels. None of the target speakers has a dialect with a distinct diphthongization and there are no clear diphthongs in the imitations. However, specific individual features, especially the pronunciation of the r-segment, are exaggerated in some of the imitations.

Some of the target speakers have a characteristic and individual speech style concerning speech rate, speech rhythm, articulation or intonation pattern. These features are easily perceived and sometimes exaggerated in the voice imitations. The impersonator also captures individual

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. 

Accepted after full paper review

characteristic features like hedges, hesitation sounds and filled pauses in these imitations.

*3.1.1. Phonetic transcription of the word 'mobilsvar'*

Three phoneticians, who listened to the voice imitations, phonetically transcribed the word *mobilsvar*. There is agreement between their transcriptions and this is shown in Table 1.

Table 1. Phonetic transcription of the word *mobilsvar*.

| | Imp | Imi-1 | Imi-2 | Imi-3 | Imi-4 | Imi-5 | Imi-6 | Imi-7 | Imi-8 | Imi-9 | Imi-10 | Imi-11 | Imi-12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| m | x | x | x | x | x | x | | x | x | x | x | x | x |
| mː | | | | | | | x | | | | | | |
| u | | x | x | | | x | x | x | | | | | |
| ɔ | x | | | x | x | | | | x | x | x | x | x |
| b | x | x | x | x | x | x | x | x | x | x | x | x | x |
| iː | x | | x | | x | | x | x | x | x | | x | x |
| ɨː | | x | | x | | x | | | | | x | | |
| l | x | x | x | x | x | x | x | x | x | x | x | x | x |
| s | x | | x | x | x | x | | | | x | | x | x |
| ʃ | | x | | | | | x | x | x | | | | |
| $s^{\theta}$ | | | | | | | | | | | x | | |
| v | x | x | x | x | x | x | x | x | x | x | x | x | x |
| ɑː | x | | x | x | x | x | x | | x | x | x | | x |
| ɑ̃ː | | x | | | | | | | | | | | |
| ɔ | | | | | | | | x | | | | x | |
| r | | | x | | x | | x | | | | | | |
| ʀ | | | | | | x | | | | | | | |
| ʁ | | x | | | | | | | x | | | | x |
| ɤ̞ | | | | | | | | | | | x | | |
| ə | | | | | | | | x | | | | | |
| ɹ | x | | | x | | | | | | x | | x | |

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. 

Accepted after full paper review

### 3.2. Acoustic analysis

#### *3.2.1. Mean F0*

There is a great variation in mean F0 (see Fig. 1), ranging from 97 to 255 Hz, between the different voice imitations. The impersonator himself speaks with a rather low mean F0 (118 Hz). Only in two of the imitations does he speak with a lower mean F0 and in both he uses a creaky voice quality. There are no great differences between the mean F0 in the whole utterance compared to mean F0 of the word *mobilsvar* in each voice imitation; this is an expected result.

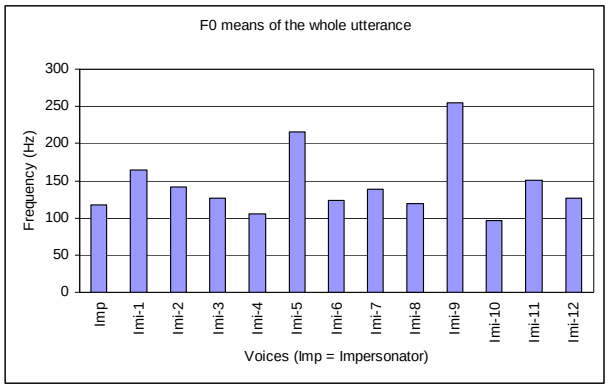


Figure 1: F0 means (in Hz) for the whole utterance.

#### *3.2.2. Duration*

The texts were different and they varied in length between 9 and 21 s. There were also duration differences in the word *mobilsvar*, which are shown in Figure 2. The duration differences are likely to depend on whether the word is emphasized or in a phrase-final position or not, and on the speech style of the voice imitation.

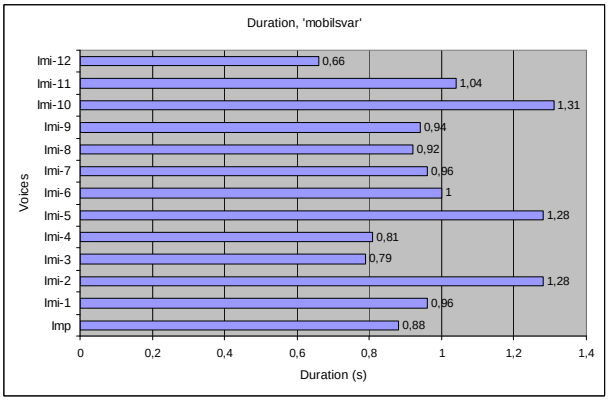


Figure 2: Duration differences in seconds (s) of the word *mobilsvar*.

The duration of the stressed vowels /i:/ and /a:/ in *mobilsvar* differs amongst the 12 different imitations, as shown in Figure 3. The individual variation of /a:/ is larger than for /i:/. The duration varies between 48 ms (Imi-3) to 293 ms (Imi-5) for the a-vowel, and between 87 ms (Imi-11) to 176 ms (Imi-10) for the i-vowel. One possible explanation for this contrast is that the a-vowels with the longest duration occur when the word *mobilsvar* is focused at the end of a phrase or is strongly emphasized. However, there is no clear pattern concerning the duration differences between the two vowels. Again, it may depend on whether the word is emphasized or in a phrase-final position or not, and the speech style of the voice imitation, as for a political speech. In the recording with the impersonator's own voice, the durations for /i:/ and /a:/ are quite close.

The 12-imitation mean duration of these two vowels is also very close to the duration for the impersonator's own vowels. The mean duration of the imitated vowel /i:/ is 0,13 s, the imitated vowel /a:/ has a mean duration of 0,15 s. For the impersonator, the duration for /i:/ is 0,14 s and for /a:/ it is 0,13 s.

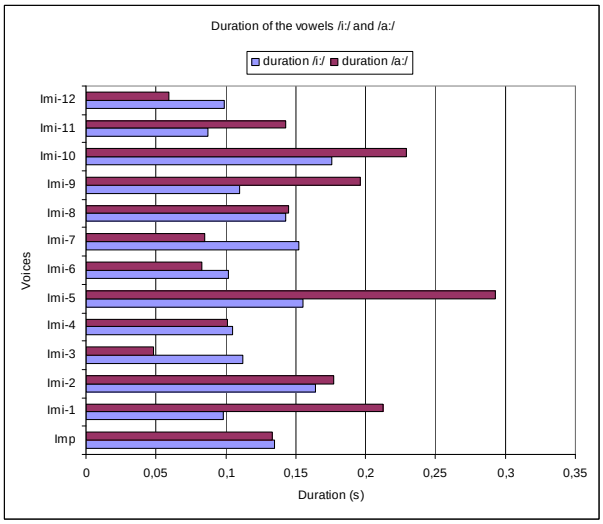


Figure 3: Duration differences in seconds (s) of the stressed vowels /i:/ and /a:/.

#### *3.2.3. Articulation rate*

The articulation rate was estimated in terms of syllables per second, in an attempt to gauge the variability exhibited by the impersonator by comparison with his different imitations. As shown in Figure 4, the impersonator has a high habitual articulation rate (5.8 syllables per second). Mean articulation rate for Swedish is about 5.0 syllables per second. Only one of the imitations, Imi-6, has a higher articulation rate. In Imi-2 and Imi-5 he speaks slowly and there are a lot of hesitation sounds and prolonged vowels.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. 

Accepted after full paper review

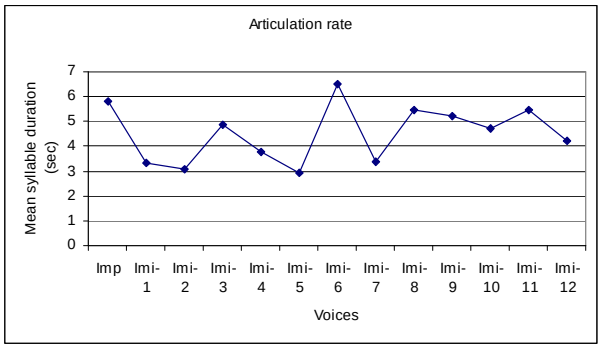


Figure 4: Articulation rate (syllables per second).

### *3.2.4. Formant frequencies*

The formant frequencies of the lowest four formants, F1-F4, of the stressed vowels /i:/ and /a:/ in the word *mobilsvar* have been compared between the different voice imitations and the impersonator's own voice. The formant frequencies were measured every 10 ms in the vowels, with the first and the last measured points about 25 ms from the beginning and the end of the vowel respectively. There is no clear pattern or measurements indicating diphthongization, and the vowels appear to be stable. Therefore, the mean value for each formant was calculated.

There are individual differences in the imitations related to the auditory impression, see the phonetic transcription, Table 1. In an attempt to uncover some connection between the impersonator's own formant frequencies and those for his imitations, both raw and averages for /i:/ and /a:/ were plotted in F1-F2 space as shown in Figure 5. The raw data reflect the range of articulatory-phonetic strategies employed by the impersonator to achieve the target voices, while the 12-imitations averages seem to successfully capture long-term characteristics of the imitator's strategies.

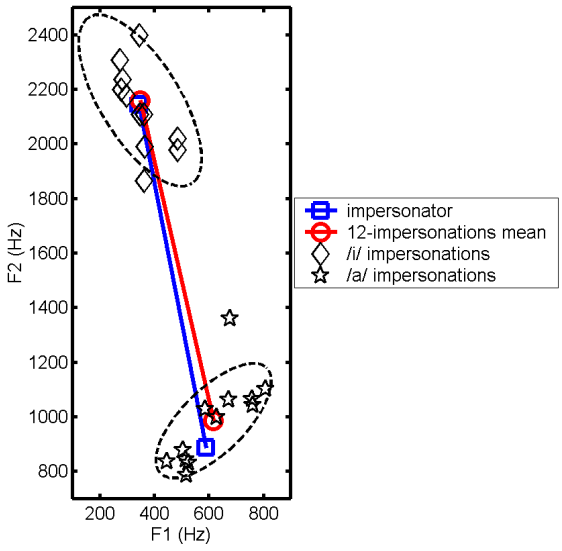


Figure 5: F1-F2 graph of the vowels /i:/ and /a:/, the impersonator, the 12 imitations and the average of the imitations.

## 4. Discussion

The listeners' general impression of these imitations is that the impersonator in question is very flexible in his voice imitations, and that he is aware of and able to imitate characteristic features of the target voices. When listening to the imitations the differences are obvious concerning both voice quality, dialect and speech style. Individual and subtle distinctions between voices are often easy to hear in an auditory analysis, but it may be hard to find acoustic correlates of these distinctions. Further research is desirable to understand more about the human perception of speech and the application in speaker identification system. An investigation of imitation success in a speaker verification system and a comparison with human perception were made (Zetterholm et al 2004). There were little agreement between listeners and the system used in this study.

In the acoustic analysis of mean F0, we have found a wide variation, which confirms the auditory impression. The impersonator's own mean F0 is 118 Hz, but in the imitations there is a variation between 97 to 255 Hz. The mean F0 in some of the imitations are related to the creaky and tense voice qualities targeted. The impersonator's own mean F0 is rather low compared to the imitated voices. Other studies of imitation (Zetterholm, 2003) show the same pattern. Could it be that it is easier to imitate a higher pitch compared to one's own without changing one's voice quality too much?

In order to understand how much duration may differ in general when one speaker reads the same text several times, one male speaker was recorded when reading the same text four times within two days at different times of the day (Zetterholm 2003). The results indicate that the within-speaker differences are rather small. However, the impersonator's data used in this study show a wide range of durations. The durational differences in the word *mobilsvar* among the imitations are obvious, and perhaps expected as argued earlier. The durations for the two stressed vowels /i:/ and /a:/ differ as well; the larger variations of /a:/ are somewhat expected given the stressed, final position of this vowel in *mobilsvar*. However, it is interesting to note that a calculation of the average of all imitations shows that the mean duration is quite close to the duration of the speech by the impersonator himself, both concerning the word *mobilsvar* (the mean duration of all imitations is 0,99 s, his own duration is 0,88 s) and the two stressed vowels. Might these results reflect the impersonator's prosodic strategies employed to achieve the target speakers.

The impersonator speaks with a high articulation rate (5.8 syllables per second) in this recording. In all but one imitation, he speaks with a lower articulation rate. In the imitations with a perceived fast speech tempo the articulation rate is 5 syllables per second, and in the imitations with a perceived slow speech tempo the articulation rate is 3-4 syllables per second. Again, this switch of speaking rate suggests another of the impersonator's articulation strategies when imitating fast and slow speech tempo. This is also in accordance to the statement that an imitation often is a stereotyping process and not an exact copy of the target speaker (Laver, 1994).

The phonetic transcription is confirmed in the measurements of the formant frequencies of the stressed vowels /i:/ and /a:/. The 12 imitations give away the range of individual strategies employed by the impersonator, while the

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. 

Accepted after full paper review

averaged data yield a long-term signature of the imitator's articulatory-phonetic strategies.

The impersonator uses words and phrases that are characteristic of the target speakers. In addition, he puts in the word *mobilsvar* in all texts. Previous studies support the hypothesis that listeners' semantic expectations would impact upon the listener's readiness to accept a voice imitation as the voice being imitated (Zetterholm et. al, 2002). However, a perception test using the same recordings as those presented in this study, indicates that it is possible to recognize voices even when listening to only one non-typical word for the target speaker (Zetterholm, 2003).

## 5. Conclusions

The results reported in this paper show that, through professional impersonation, it is possible to identify and to imitate another speaker's voice and speech behaviour with general success, according to comments from our panel of listeners.

The acoustic analyses of the imitations made by our impersonator highlight the flexibility of his voice and his ability to change his habitual settings. The results obtained from the average of the 12 imitations, both concerning the duration and the formant frequencies, suggest that there are short-and long-term strategies employed by the impersonator. The elucidation of these strategies will require further study. The findings concerning articulation rate, and its correlations with fast and slow speech tempo, might give us insights into pattern and processes in the imitation task.

Further work in these areas will ultimately contribute towards uncovering more robust ways to approach the problem of forensic voice identification.

## 6. Acknowledgements

This study is partly funded by a grant from the Bank of Swedish Tercentenary Foundation Dnr K2002-1121:1-4 to Umeå University, Sweden, for the project 'Imitated voices: A research project with applications for security and the law'. Many thanks to Dr Frantz Clermont for fruitful discussions and the F1-F2 graph in Figure 5.

## 7. References

Bruce, G. (1970). Diphthongization in the Malmö dialect. Lund University, *Working Papers 3*: 1-19.

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

Elert, C-C. (1991). *Allmän och svensk fonetik*. Stockholm: Norstedts Förlag AB.

Elliott, J. (2000). Auditory and F-pattern variations in Australian okay: a forensic investigation. *SST-2000: The Eight Australian International Conference on Speech Science and Technology*, Canberra, Australia.

Laver, J. (1994). *Principles of phonetics*. Cambridge, Cambridge University Press.

Pittam, J. (1994). *Voice in Social Interaction; An Interdisciplinary Approach*. Thousand Oaks, SAGE Publications.

Rose, P. (1999). Differences and distinguishability in the acoustic characteristics of hello in voices of similar-sounding speakers – a forensic phonetic investigation. *Australian Review of Applied Linguistics 21 (2)*: 1-42.

Scherer, K.R. (1979). Personality markers in speech. In K.R. Scherer and H. Giles (eds.) *Social markers in speech*. Cambridge, Cambridge University Press: 147-209.

Schlichting, F. and K.P.H. Sullivan (1997) The imitated voice – a problem for voice line-ups? *Forensic Linguistics 4 (1)*: 148-165.

Zetterholm, E., K.P.H. Sullivan and J. van Doorn (2002). The Impact of Semantic Expectation on the Acceptance of a Voice Imitation. *Proceedings of the 9th Australian International Conference on Speech Science & Technology*, Melbourne, Dec 2002: 379-384.

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.

Zetterholm, E., M. Blomberg and D. Elenius (2004). A comparison between human perception and a speaker verification system score of a voice imitation. *Proceedings of the 10th Australian International Conference on Speech Science & Technology*, Sydney, Dec 2004: 393-397.

Zetterholm, E. (2006). (in print) Detection of speaker characteristics using voice imitation. In C. Müller and S. Schötz (eds.) *Speaker Classification*. Springer LNCS/LNAI series.