HOW FLEXIBLE IS THE HUMAN VOICE? – A CASE STUDY OF MIMICRY.

Anders Eriksson and Pär Wretling

Department of Phonetics

Umeå University, S-901 87 Umeå, Sweden

E-mail: anderse@ling.umu.se and wretling@ling.umu.se

ABSTRACT

The investigation presented here is a case study of mimicry in which a professional impersonation artist imitated three well-known Swedish public figures. The speech material consisted of recorded material taped from radio/TV shows, imitations of these speeches in which the artist tried to mimic the speeches as closely as possible, and the same speech material recorded with the artist using his own natural voice. The aim of the study was to investigate how closely the imitations matched selected acoustic parameters of the original recordings. It was found that he was able to mimic global speech rate very closely, but timing at the segmental level showed little or no change in the direction of the targets. Mean fundamental frequency and variation matched the targets very closely. Target formant frequencies were attained with varying success. For two of the three target voices the vowel space of the imitation was intermediate between that of the artist's own voice and the target. In the third case there was no apparent reduction in distance. With respect to individual vowels it was generally, but not always, the case that the formant frequencies of the mimicked vowels were closer to the original than those of the artist's own voice.

1. INTRODUCTION

Mimicry has been studied phonetically to a very minor extent. This is quite surprising considering the interest in the phenomenon among the general public. The phenomenon has, however, an interest that goes far beyond the area of public entertainment. From a purely phonetic point of view it may tell us a lot about the flexibility of the human voice – to what extent, and in what ways, is it possible to modulate one's voice. The phenomenon should also be of great relevance in forensic phonetic research. Although it may be true, as claimed in the literature [1], that there are few actual cases of fraud by mimicry in the courts today, this situation may change drastically in the very near future if automatic voice recognition is used as a means of personal identification in security systems. The performance of a professional impersonation artist probably represents state-of-the-art performance in this area and material of that kind should, therefore, represent interesting test cases.

2. BACKGROUND

Only a handful of studies exist in this area, and to our knowledge only two that deal with acoustic parameters in some detail. Bessler [2] has studied a caricatured impersonation of de Gaulle. The most relevant finding with respect to the present study was the fact that the impersonator exaggerated both mean fundamental frequency level and range. Stressed syllable durations were also exaggerated. The impersonation was primarily meant for entertainment purposes and not accuracy of imitation. The generalizability of the results may therefore be questioned. In the other study [3], vowel formant frequencies and fundamental frequencies in imitations were compared with the corresponding values for the original voices. Although the imitators managed to change their formant and fundamental frequencies in the direction of the target values "they were not able to adapt these parameters to match or even be similar to those of the imitated persons." [3, p. 1842]

3. METHOD

3.1 Speech material used in the study

The target material consisted of approximately 30 s long excerpts of uninterrupted speech by three well-known Swedish public figures recorded from radio/TV shows. Recordings were then made of a professional Swedish impersonation artist trying to imitate the speakers in these excerpts as closely as possible. These imitations were not intended to be entertaining, but explicitly meant to mimic the target voices and speech styles as closely as possible. In particular the impersonator kept a close eye on timing, trying to keep speech rates the same as those of the targets. The same speech material was also recorded with the artist using his own natural voice and speaking style. This was done in order to evaluate how much the artist had to change his natural voice in order to reach the targets.

3.2 Acoustic analyses

The speech material was digitized at 16kHz/16bps. *ESPS/waves+*™ routines were used in all subsequent

acoustic analyses. All speech files were labelled at both word and segment level. Fundamental frequency and formant frequencies were then computed. First four formants were calculated for all Swedish vowel allophones which were present in the material. All tokens were numbered consecutively to ensure that comparisons between tokens in the original, imitation and the artist's own voice were made between vowels in identical contexts.

4. RESULTS

4.1 Timing

The impersonator succeeded in mimicking global speech rate very closely as shown in Table 1. It has to be noted, however, that global speaking rates for the target speakers were very similar to the impersonator's own preferred speaking rate except for target voice 'Loket'.

Table 1. Total durations (seconds) for the three sets of speech samples and the differences between imitations and the Target. 'Natural' refers to the renditions by the impersonator in his own natural speaking style. 'Bildt', 'Loket', and 'Parnevik' are the three target voices.

	Bildt		***Loket***		***Parnevik***	
	Total	**Diff.**	**Total**	**Diff.**	**Total**	**Diff.**
Target	30.44		43.20		29.90	
Imitation	30.89	0.45	44.59	1.39	31.46	1.56
Natural	30.39	–0.05	50.02	6.82	31.40	1.50

A more detailed analysis of timing showed, however, that local deviations were almost on the same order as the differences in global durations. This may be clearly seen in Figure 1.

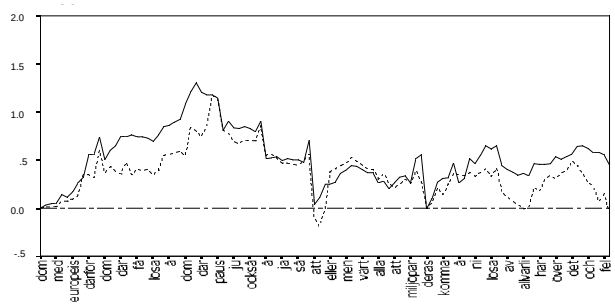

Figure 1. Deviation (seconds) at the beginning of each consecutive word in the imitation (solid line) and the impersonator's natural rendition (dotted line) from the corresponding time points in the target (Bildt).

For this particular imitation, the difference in global timing was only 0.43 s, whereas the greatest local deviation was almost 1.5 s. It is also worth noting that deviations for the imitation and the natural rendition are almost perfectly parallel. The impersonator is, thus, able to control global timing well enough to 'arrive' at the end of the sentence almost perfectly on time. However, he seems to have less control over articulatory timing at a more local level. This impression is confirmed by a closer inspection of durations at the word level. Word durations were closer to the target durations than to corresponding durations in the natural versions in only 50%, 40%, and 33% of the cases in the three imitations respectively. This means that word durations in the imitations were more similar to the impersonator's own speaking style than to the target in two of the three cases and somewhere in between in the third case.

As may be seen in Table 2, word durations in the imitations correlate better with the impersonator's natural versions than with the targets. The difference is not striking and all correlations are significant, but there is the same trend in all three versions towards better agreement between the imitation and the natural version.

Table 2. Correlation coefficients for word durations.

	Bildt	***Loket***	***Parnevik***
Target – Imitation	0.83	0.79	0.59
Natural – Imitation	0.90	0.81	0.82

A study of timing at the segmental level confirms the greater similarity between the natural versions and the imitations than between the imitations and the targets. As may be seen in Figure 2, this may be true even in the case where the total duration of a phrase is almost identical in the three versions.

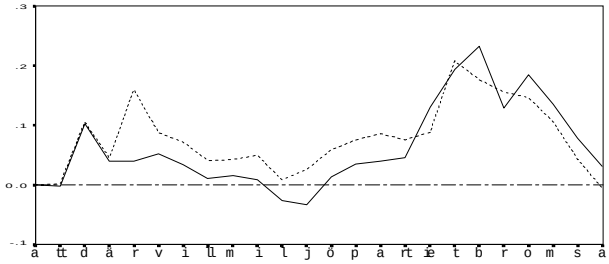

Figure 2. Deviation (seconds) at the beginning of each consecutive segment in the imitation (solid line) and the impersonator's natural rendition (dotted line) from the corresponding time points in the target (Bildt) in a typical phrase.

As is the case for word durations in the entire speech, segment timing deviations in this case are almost parallel for the imitation and the natural rendition indicating that articulatory timing patterns are very similar between the two and different from that in the target. The pattern shown in Figure 2 is typical of all phrases studied so far. Articulatory timing at the segmental level changes very little in the imitations. Even when speech rate changes globally, proportions between segment durations seem to remain basically the same.

4.2 Fundamental frequency

As was the case with timing, the impersonator succeeded in attaining the global targets very well. (See Table 3!) For target voice 'Bildt', mean fundamental frequency is exactly the same as that of the target. For the other two

target voices, the means deviate, but only marginally (4 Hz in both cases). It should be noted that in two of the cases the impersonator had to increase his mean fundamental frequency by about 15 Hz and in the third case lower it by 30 Hz, but modifications were equally successful in both directions.

Global variation, here expressed as the standard deviation in semitones, is modified in the direction of the target voice in all three imitations, but without quite reaching the same degree of variation in any of the cases.

Table 3. Mean fundamental frequency (Hz) and standard deviation (semitones) for the entire utterances.

	Bildt		*Loket*		*Parnevik*	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Target	134	4.11	146	3.41	85	3.76
Imitation	134	3.49	142	3.23	89	3.63
Natural	129	3.45	130	2.68	119	2.89

The material used here has so far not been analysed in any detail with respect to local variation, primarily prosody. Preliminary analyses suggest that the impersonator succeeds in imitating prosodic contours quite well, but a quantitative evaluation has yet to be made.

4.3 Formant frequencies

The impersonator's vowel space was greater than that of the target voice in all cases. Globally, he therefore had to centralise his vowels to approach the target voice formants. He managed this with varying success. The most successful case is shown in Figure 3. Here the vowel space of the imitation is roughly intermediate between the impersonator's natural space and the target space (Bildt). In the imitation of target voice 'Parnevik' he was also quite successful, globally, whereas for the target voice 'Loket' he did not manage to adjust his vowel space to any appreciable extent.

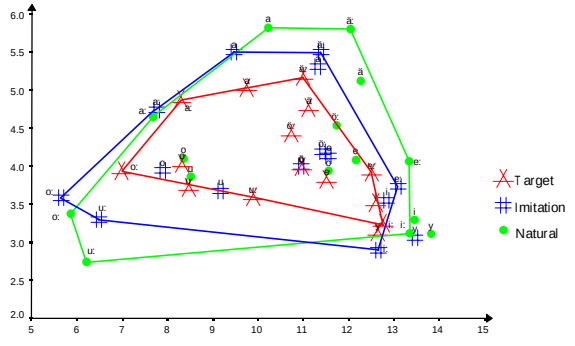

Figure 3. Vowel spaces for one of the targets (Bildt), the corresponding imitation, and the impersonator's natural voice. Abscissa is F_2 and ordinate F_1 (Bark scales).

If Euclidean distance in vowel space is used as a distance measure, we may express mean distance between the vowels in 2-dimensional vowel space numerically. Table 4 shows these distances. As may be seen, distances decrease in two of the three cases. Interestingly, the imitation that falls out of line ('Loket') is the one that the impersonator himself regarded as the least successful.

Table 4. Mean Euclidean distances (in Bark) in 2-dimensional vowel space (mean values based on the median values for the individual vowels).

	Bildt	*Loket*	*Parnevik*
Target – Imitation	0.75	0.87	0.54
Natural – Imitation	1.01	0.81	0.93

If mean formant values (Table 5) are taken to be representative of the vocal tract characteristics of a speaker, it may be seen, however, that the impersonator does not differ very much in this respect from the targets. It is, therefore, unlikely that failure by the impersonator to hit target formant values should be due to physiological limitations.

Table 5. Mean formant frequencies (in Hz) for the three sets of speech samples.

	Bildt			*Loket*			*Parnevik*		
	F_1	**F_2**	**F_3**	**F_1**	**F_2**	**F_3**	**F_1**	**F_2**	**F_3**
Target	398	1424	2536	483	1464	2586	408	1437	2823
Imitation	390	1461	2684	447	1515	2720	385	1273	2481
Natural	395	1563	2671	447	1547	2614	416	1430	2580

A detailed analysis of the different vowel types showed considerable variation with respect to how well the impersonator succeeded in approaching the formant values of the targets. The Euclidean distance in Bark was used as the distance measure to describe these deviations. Figure 4 shows, in graphical form, the result of one of these analyses.

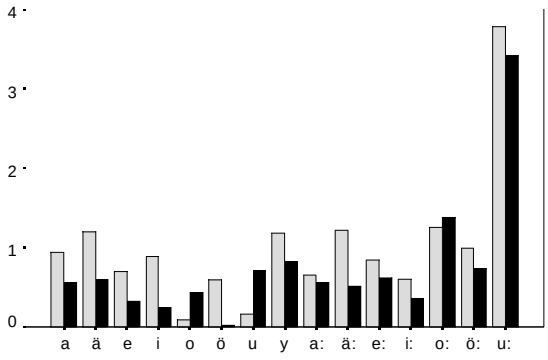

Figure 4. Distances in Bark, in 2-dimensional vowel space, between the impersonator's natural voice and the target Bildt (grey), and the imitation and the target (black).

As may be seen in the diagram most, but not all, distances decrease in the imitation. The imitation on which Figure 4 is based represent the most successful case (Bildt). In the other imitations, target values were

reached less successfully. Many of these differences should be clearly audible.

5. CONCLUSIONS

5.1 Timing

The results obtained here indicate that a speaker may vary speaking rate globally to produce a phrase or longer stretches of speech within a given time limit. Articulatory timing patterns at a more local level appear to be more rigid. For a given speech rate, durations of individual words changed little for this speaker and durations correlated more closely with the impersonator's natural voice than with the target. At the segmental level patterns seem to be even more rigid. What this suggests is that timing patterns at the segmental level may be very difficult to change. A speaker may change his or her speaking rate globally, but the relative durations of the segments remain fairly invariant for a given speech rate. This is all the more surprising since in this case the speaker consciously tried *not* to speak in his own natural speaking style but to adopt the style of another speaker. From a forensic point of view this may be highly relevant. If this turns out to hold true for speakers in general, it would be an indication that segmental timing may be more or less hard coded in an individual speaker. If this is the case, timing may be a valuable criterion in speaker identification.

5.2 Fundamental frequency

Fundamental frequency changes were more successful than changes in timing. The impersonator came very close to the global targets. Global means, thus, do not seem to be of much use for automatic speaker identification. Perception tests [3] indicate that these properties do not play a great role in the identification of speakers by listeners either. For the data presented here, it is unlikely that any of the global differences are even noticeable in perception, with the possible exception of the difference in standard deviation for target Bildt (0.62 s.t.). In a perception study [4] where global variation in semitones was varied in 0.85 s.t. steps, each step made a clearly audible difference. Whether this is any help in identifying a particular speaker is, of course, an entirely different matter.

It seems reasonable to suggest, that at least part of the explanation for the more successful fundamental frequency modifications, is that tonal modifications are simply easier to achieve because we are used to constantly modifying the 'speech melody'. But it is also likely that fundamental frequency contours are what impersonators find most rewarding to imitate since they express such characteristic properties as prosody and paralinguistic variation and are also quite revealing with respect to dialectal variation.

5.3 Formant frequencies

The results lend support to the findings by Endres *et al*. [3] that it is difficult to accurately modify formant frequencies towards a given target. At least this seems to be true in ordinary mimicry. A caveat is in place though – close mimicking of spectral qualities may not be part of mimicry in the form it is used for entertainment. And although it is true that the imitations studied here were not meant for entertainment purposes, it may not be ruled out that the tradition in which the subject was trained may be at least a partial explanation for the poorer results in this respect. It, therefore, remains an open question whether the performance could be significantly improved if the impersonator were to consciously and systematically practice this particular aspect of mimicry.

From the point of view of automatic speaker recognition, the results confirm the importance of the spectral dimension. Even a highly trained person does not succeed in hitting target values for vowels formants with any great precision. To be able to evaluate the results more precisely, however, one would have to compare the differences between the imitations and the targets with the naturally occurring variation in the target voices themselves. If one should find that the deviations from the target produced by a skilled impersonator lie within the range of normal variation for a speaker, this would, of course, weaken the reliability of the spectral cues.

With respect to the question posed in the title of this paper, the results obtained here indicate that timing at the segmental level is the least flexible aspect of speech production, that is the aspect which a speaker is least able to modify in a desired direction. What a speaker may vary reasonably freely is global speaking rate, but once that is done segment durations seem to be determined within rather narrow limits. Speakers seem to be much better at varying formant frequencies in the direction of a given target, but the precision with which they are able to hit the targets varies a great deal and is usually not good enough for the vowels to be perceptually indistinguishable from the targets.

6. REFERENCES

[1] Künzel, H. J. (1987) *Sprechererkennung: Grundzüge forensicher Sprachverarbeitung*. Heidelberg: Kriminalistik-Verlag.

[2] Bessler, P. (1991) La caricature de de Gaulle par Tissot: Étude phonostylistique. In: *Information/Communication*, **12**, 19–32. Canadian Scholars' Press.

[3] Endres, W., W. Bambach & G. Flösser. (1971) Voice spectrograms as a function of age, voice disguise, and voice imitation. *J. Acoust. Soc. Am.*, **49**, 1842–1848.

[4] Traunmüller, H. & A. Eriksson. (1995) The perceptual evaluation of F_0 excursions in speech as evidenced in liveliness estimations. *J. Acoust. Soc. Am.*, **97**, 1905–1915.