Umeå University, Department of Philosophy and Linguistics
PHONUM 9 (2003), *161-164*
Available online at http://www.ling.umu.se/fonetik2003/

Why Foil 4? A first look

Erik Eriksson[1], Frank Kügler[2], Kirk P. H. Sullivan[1,3], Jan van Doorn[4] and Elisabeth Zetterholm[5]

[1]Department of Philosophy and Linguistics, Umeå University, [2]Department of Linguistics, Potsdam University, [3]Department of Computing Science, Umeå University, [4]Department of Clinical Sciences, Umeå University, [5]Department of Linguistics and Phonetics, Lund University

Recent studies have shown that ability to recognize an imitated voice is affected by topic and familiarity with the person being imitated. The signal detection methods used in those studies give no specific information about the distribution of non matching positive responses (false alarms) amongst voices used as distracters (foils). This paper shows that there was a selective listener preference for Foil 4. However, preliminary comparison of Foil 4 with the target imitated voice has not revealed conclusively any similarities in speech features that account for listeners' preferential selection.

1. Introduction

The degree to which listeners can be misled by voice imitation has been investigated in a series of recent studies (e.g. Schlichting and Sullivan, 1997; Zetterholm, Sullivan and van Doorn, 2002, Zetterholm et al 2002 and Zetterholm, 2003). Recently (Zetterholm et al, submitted) the data reported in Zetterholm et al (2002), Sullivan et al (2002), and Zetterholm et al (2003) were pooled and reanalyzed. For the reanalysis the variable "self-reported familiarity with the target voice" was introduced. For both the group who reported that they were familiar with the target voice and the group who reported that they were not, the distribution of the positive voice selections was examined. This revealed that foil, or distracter, 4 was being positively selected more frequently than the other three foils or the imitator's natural voice. The aim of this study is to discover the reasons for this selection preference and consider the implications for voice witness evidence.

2. The perception tests

The data that was reanalyzed in Zetterholm et al (submitted) was collected using the experimental method that is summarized below.

2.1. The voices and the participants

The voices consisted of a set of seven recordings of the same text of a political speech, and one recording of a non-political topic (how to bake a cake). The political recordings were an original by Carl Bildt (PS-Bildt), a professional imitation of the voice of Carl Bildt (AM1), the natural voice of the professional imitator (PS-AM) and four other male voices, referred to hereafter as 'foils'. The non-political recording was a free voice imitation of Carl Bildt explaining how to bake a cake (AMK). Two of the recordings (AM1 and AMK) were used as familiarization passages, and the remaining six were used as the basis for the voice line-up. Table 1 reports information about these foils and the duration and F0 of the recordings.

The listeners were all randomly selected and no listener reported any hearing damage. The listeners were native Swedish speakers from Umeå, Örebro and Lund. The listeners from each of these towns were divided into two approximately equal sub-groups. After the perception experiments, all the participants reported whether they were Familiar with the voice of Carl Bildt (FCB+) or not (FCB-), and if they were familiar with any of the other voices used in the experiment or not. The descriptive statistics of the listener groups is reported in Table 2.

2.2. The experiment

Two experiments were constructed; each began with a familiarization voice. The experiments differed only in the familiarization passage. Experiment 1 used the political passage imitation (AM1), while Experiment 2 used the "how to bake a cake" passage imitation (AMK).

For both experimental set-ups, the line-up was constructed from the six recordings of the political passage that were not used for the familiarization task i.e. recordings by PS-Bildt, PS-AM, and the four other male voices. Three separate segments were spliced out from each of the six recordings. Each segment was repeated three times in the line-up, giving a total of 54 speech stimuli in the line-up (3 repetitions x 3 speech samples x 6 speakers). The line-up voices thus contained PS-Bildt, PS-AM and foils as the test voices.

The two experiments were conducted separately. Both sub-groups were first familiarized with the voice of the target speaker they were to identify, and were told that they would be asked to recognize it later. Then they listened to a CD containing the 54 speech stimuli, presented in a randomized order. The listeners were instructed to respond 'Yes' on a response sheet whenever they recognized the voice from the familiarization recording, and 'No' when they did not.

Table 1: The voices, their age and dialect background, and each recording's duration, F0 mean and standard deviation.

Voice	Age	Dialect Background	Duration	F0 mean	Std dev F0
PS-Bildt	ca45	Halland	31 sec	135 Hz	35 Hz
AM1			36 sec	139 Hz	30 Hz
PS-AM	ca40	Västergötland	34 sec	123 Hz	35 Hz
AMK			37 sec	123 Hz	35 Hz
Foil 1	28	Jämtland	32 sec	132 Hz	59 Hz
Foil 2	20	Västerbotten	38 sec	134 Hz	36 Hz
Foil 3	22	Halland	29 sec	137 Hz	19 Hz
Foil 4	54	Skåne	34 sec	118 Hz	46 Hz

Table 2: The descriptive statistics of the listener groups. Missing gender data points are not explicitly given. Exp 1 heard the AM1 training passage; Exp2 heard the AMK training passage

Exp	No	No. Male	No. Female	Mean Age	SD Age
1	130	51	78	25.30	12.49
2	124	52	68	26.77	13.13

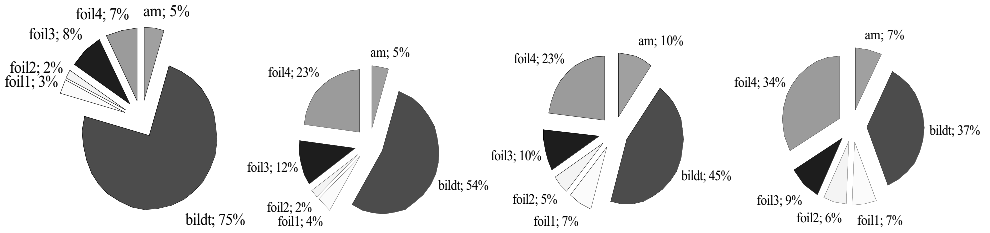

Figure 1: The distribution of yes responses. The chart to the left shows the distribution for the FCB+ listeners who heard the training passage AM1 (experiment 1); the chart second to the left shows the distribution for the FCB+ listeners who heard the training passage AMK (experiment 2); the chart second to the right shows the distribution for the FCB- listeners who heard the training passage AM1 (experiment 1); the chart to the right shows the distribution for the FCB- listeners who heard the training passage AMK (experiment 2).

2.3. Selection distribution

Figure 1 shows the distribution of yes responses. The number of voices in the line-up sets the upper limit for the number yes responses a participant can give. Most listeners have a response bias towards one response possibility and the number of responses will fall between zero and the upper limit, inclusive. Further, few listeners are able to perfectly select the target voice and reject all other voices. For this reason, the papers that investigated the impact of the different training texts, e.g. Zetterholm, Sullivan and van Doorn, (2002) calculated values for d’ and c (see Green and Sweets, 1966), rather than examine just the yes responses. However, an examination of the yes selection distribution provides an insight into which foil(s) distracted best, in this case foil 4. The following acoustic and auditive analysis investigates the listeners’ preference for foil 4 and avoidance of the natural voice of the imitator. Differences and similarities between the line-up voices were sought using mean F0 of the whole passage, and phonetic features of the selected phrases.

3. A preliminary acoustic and auditive analysis of the voices

The mean F0 values, measured using Praat, are reported in Table 1 along with the standard variation in F0 for each recording. Carl Bildt uses a uvular trill [ʀ], this is reproduced by AM in his imitations. However, Foils 3 and 4 use the back fricative [ʁ] and Foils 1 and 2 an alveolar tremulant [r]. In general AM successfully copied Carl Bildt’s use of long intonation phrases, his tendency to draw out voiced segments such as the voiced stop [b] and vowels, and the suggestion of creaky voice in both imitations. However, AM failed to successfully impersonate Bildt’s feature trait of cutting the end of the last syllable. Foil 1 articulates clearly and does not reduce the word ‘det’ in the line-up phrase ‘det är illa och det är fel’. In the Swedish word ‘kanske’ (eng: *perhaps*) he uses the [ʂ] sje-sound, that is a dialect marker for nothern Swedish. Foil 2 gives the auditive impression of having a high pitched voice. He articulates clearly, yet has many reductions. He uses the intonation, vowels and [ʂ] sje-sound of a northern Swedish dialect. Foil 3 has a clear Halland accent. He talks quickly, yet clearly and uses the sje-sound [ɧ] in the word ‘kanske’; this is to be expected for someone coming from this area. Foil 4 has a clear Scanian accent, yet with few diphthongs. He uses a uvular fricative -r [ʁ] and the sje-sound [ɧ].

4. Conclusion

Explanation of preference for Foil 4 by the listeners is not easily explained by mean F0 of the whole passage, although it may explain why the effect was greater for the kitchen imitation listeners (where mean F0s were closer together) than the PS imitation (where they were furthest apart). Neither is the preference easily explained by the most prominent phonetic features in the individual phrases of each foil voice. The use of a back r-sound may be one factor even if Carl Bildt uses a tremulant and foil 4 a fricative r-sound, but this is not clear either. Foil 3 is dialectally nearest in his imitation, yet is less successful that foil 4.

The age of the speaker could have the strongest explanatory power. Yet, in their paper examining the estimation of speaker age across languages, Braun and Cerrato (1999) reviewed the literature in this area and reported no unambiguous listener ability here. This is also the situation in this study; the impact of age is unclear. Foil 4 is the foil nearest in age to Carl Bildt, but AM is nearer in age and his natural voice was not selected as frequently. Further, foil 3 was selected more frequently than AM, even though he was much younger.

A more detailed examination of the stimuli is required. This will examine the mean F0, the F0 contours of the selected phrases, accent realization and look carefully at the relative timing of these phrases.

5. Acknowledgement

This paper was supported by The Bank for Sweden Tercentenary Foundation through their funding of the project “Imitated voices: A research project with applications for security and the law” Dnr K2002-1121:1, Dnr K2002-1121:2.

6. References

Braun, A. and Cerrato, L. (1999) Estimating speaker age across languages. *In Proceedings of ICPhS99*, pp.1369-1372.

Green, D.M. and Sweets, J.A. (1966) *Signal Detection Theory and Psychophysics*, New York: Wiley.

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

Sullivan, K. P. H., Zetterholm, E., van Doorn, J., Green, J., Kügler, F. and Eriksson, E. (2002) The effect of removing semantic information upon the impact of a voice imitation. *In SST2002 Proc*eedings, Melbourne, Australia, pp. 291-296.

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

Zetterholm, E., Sullivan, K.P.H. and van Doorn, J. (2002) The impact of semantic expectation on the acceptance of a voice imitation. *In SST2002 Proceedings*, Melbourne, Australia, pp. 379–384.

Zetterholm, E., Sullivan, K. P. H., Green, J., Eriksson, E. and Czigler, P. E. (In press) “Imitation, expectation and acceptance: The role of age and first langauge in a nordic setting. *In ICPhS 2003 Proceedings*.

Zetterholm, E., Sullivan, K. P. H., Green, J., Eriksson, E., van Doorn, J, and Czigler, P.E. (Submitted to Eurospeech 2003) Who knows Carl Bildt — And what if you don’t?