This assumption is strengthened by the subjects' reaction to the other two features investigated. Both intonation and creaky voice have the capacity to signal an upcoming boundary so that they are more likely to facilitate the detection of a hesitation in a phrase-internal position, where a boundary is unexpected, than between two grammatical phrases. This dependence on syntax is not unexpected: vast numbers of production studies have shown the strength of prosodic signalling to depend on the strength of the syntactic boundary.

In conclusion, our results indicate that the perception of hesitation is strongly influenced by deviations from an expected temporal pattern. In addition, different syntactic conditions have an effect on how much changes in prosodic features like the F0 contour and retardation and the presence of creaky voice contribute to the perception of hesitation. In view of this, the modelling of hesitation in speech technology applications should take account of the supporting roles that F0 and creak can play in achieving a realistic impression of hesitation.

An important step in the modelling of spontaneous speech would be to include predictions of different degrees of hesitations depending on the utterance structure. To do this, data are required of the distribution of hesitations, see e.g. Strangert (2004). Our long-term goal is to build a synthesis model which is able to produce spontaneous speech on the basis of such data. An even more long-term goal is to include other kinds of disfluencies as well, and to integrate the model in a conversational dialogue system, cf. Callaway (2003).

Acknowledgements

We thank Jens Edlund, CTT, for designing the test environment, and Thierry Deschamps, Umeå University, for technical support in performing the experiments. This work was supported by The Swedish Research Council (VR) and The Swedish Agency for Innovation Systems (VINNOVA).

References

Callaway, C., 2003. Do we need deep generation of disfluent dialogue? In *AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue, Tech. Rep. SS-03-07*. Menlo Park, CA: AAAI Press.

Carlson, R. & B. Granström, 1997. Speech synthesis. In W.J. Hardcastle & J. Laver (eds.), *The Handbook of Phonetic Science*. Oxford: Blackwell Publ., 768-788.

Carlson, R., K. Gustafson & E. Strangert, 2006. Modelling Hesitation for Synthesis of Spontaneous Speech. *Proc. Speech Prosody 2006*, Dresden.

Eklund, R., 2004. *Disfluency in Swedish human-human and human-machine travel booking dialogues*. Dissertation 882, Linköping Studies in Science and Technology.

Horne, M., J. Frid, B. Lastow, G. Bruce & A. Svensson, 2003. Hesitation disfluencies in Swedish: Prosodic and segmental correlates. *Proc. 15th ICPhS*, Barcelona, 2429-2432.

Klatt, D. & L. Klatt, 1990. Analysis, synthesis and perception of voice quality variations among female and male talkers. *JASA 87*, 820-857.

Lövgren, T. & J. van Doorn, 2005. Influence of manipulation of short silent pause duration on speech fluency. *Proc. DISS2005*, 123-126.

Strangert, E., 2004. Speech chunks in conversation: Syntactic and prosodic aspects. *Proc. Speech Prosody 2004*, Nara, 305-308.

Strangert, E. & R. Carlson, 2006. On modelling and synthesis of conversational speech. *Proc. Nordic Prosody IX, 2004*, Lund, 255-264.

Sundaram, S. & S. Narayanan, 2003. An empirical text transformation method for spontaneous speech synthesizers. *Proc. Interspeech 2003*, Geneva.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 25–28

F-pattern Analysis of Professional Imitations of "hallå" in three Swedish Dialects

Frantz Clermont and Elisabeth Zetterholm
Dept. of Linguistics and Phonetics, Centre for Languages and Literature, Lund University
{frantz.clermont|elisabeth.zetterholm}@ling.lu.se

Abstract

We describe preliminary results of an acoustic-phonetic study of voice imitations, which is ultimately aimed towards developing an explanatory approach to similar-sounding voices. Such voices are readily obtained by way of imitations, which were elicited by asking an adult-male, professional imitator to utter two tokens of the Swedish word "hallå" in a telephone-answering situation and three Swedish dialects (Gothenburg, Stockholm, Skania). Formant-frequency (F1, F2, F3, F4) patterns were measured at several landmarks of the main phonetic segments ('a', 'l', 'å'), and cross-examined using the imitator's token-averaged F-pattern and those obtained by imitation. The final 'å'-segment seems to carry the bulk of differences across imitations, and between the imitator's patterns and those of his imitations. There is however a notable constancy in F1 and F2 from the 'a'-segment nearly to the end of the 'l'-segment, where the imitator seems to have had fewer degrees of articulatory freedom.

1 Introduction

It is an interesting fact but all the same a challenging one in forensic voice identification, that certain voices should sound similar (Rose & Duncan, 1995), even though they originate from different persons with differing vocal-tract structures and speaking habits. It is also a familiar observation (Zetterholm, 2003) that human listeners can associate an imitated voice with the imitated person. However, there are no definite explanations for similar-sounding voices, and thus there is still no definite approach for understanding their confusability. Nor are there any systematic insights into the degree of success that is achievable in trying to identify an imitator's voice from his/her imitations. Some valiant attempts have been made in the past to characterise the effects of disguise on voice identification by human listeners. More recently, there have been some useful efforts to evaluate the robustness of speaker identification systems (Zetterholm et al., 2005). The results are however consistent in that "it is possible to trick both human listeners and a speaker verification system" (Zetterholm et al., 2005: p. 254), and that there are still no clear explanations.

Overall, the knowledge landscape around the issue of similarity of voices appears to be quite sparse, yet this issue is at the core of the problem of voice identification, which has grown pressing in dealing with forensic-phonetic evaluation of legal and security cases. Our ultimate objective, therefore, is to use acoustic, articulatory and perceptual manifestations of imitated voices as pathways for developing a more explanatory approach to similar-sounding voices than available to date.

The present study describes a preliminary step in the acoustic-phonetic analysis of imitations of the word "hallå" in three dialects of Swedish. The formant-frequency patterns obtained are enlightening from a phenomenological and a methodological point of view.

2 Imitations of the Swedish word "hallå" – the speech material

The material gathered thus far consists of auditorily-validated imitations of the Swedish word "hallå". An adult-male, professional imitator was asked to first produce the word in his own usual way. The imitator is a long-term resident of an area close to Gothenburg and, therefore, his speaking habits are presumed to carry some characteristics of the Gothenburg dialect. He was asked to also produce imitations of "hallå" in situations such as: (i) answering the telephone, (ii) signalling arrival at home, and (iii) greeting a long-lost friend, all in 5 Swedish dialects (Gothenburg, Stockholm, Skania, Småland, Norrland). The 2 tokens obtained for the first 3 dialects in situation (i) were retained for this preliminary study. The recordings took place in the anechoic chamber recently built at Lund University. The analogue signals were sampled at 44 kHz, and then down-sampled by a factor of 4 for formant-frequency analyses.

3 Formant-frequency parameterisation

3.1 Formant-tracking procedure

The voiced region of every waveform was isolated using a spectrographic representation, concurrently with auditory validation. Formants were estimated using Linear-Prediction (LP) analyses through Hanning-windowed frames of 30-msec duration, by steps of 10 msecs, and a pre-emphasis of 0.98. For 25% of the data used for this study, the LP-order had to be increased to 18 from a default value of 14. For each voiced interval, the LP-analyses yielded a set of frame-by-frame poles, among which F1, F2, F3 and F4 were estimated using a method (Clermont, 1992) based on cepstral analysis-by-synthesis and dynamic programming.

3.2 Landmark selection along the time axis

The expectedly-varying durations amongst the "hallå" tokens raise the non-trivial problem of mapping their F-patterns onto a common time base. We sought a solution to this problem by looking at the relative durations of the main phonetic segments ('a', 'l', 'å'), which were demarcated manually. The token-averaged durations for imitated and imitator's segments are superimposed in Fig. 1, together with the overall mean per segment.

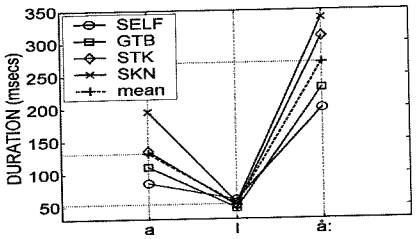

Figure 1. Segmental durations: Mean ratio of ~3 to 1 for 'a', ~5 to 1 for 'å', relative to 'l'.

Interestingly, the durations for the imitator's 'a'- and 'å'-segments are closer to those measured for his Gothenburg imitations, and smaller than those measured for his Skanian and Stockholm imitations. Fig. 1 also indicates that the medial 'l'-segment has a duration that is tightly clustered around 50 msecs and, therefore, it is a suitable reference to which the other segments can be related. On the average, the duration ratio relative to the 'l'-segment is about 3 to 1 for 'a', and 5 to 1 for 'å'. A total of 45 landmarks were thus selected such that, if 5 are arbitrarily allocated for the 'l'-segment, there are 3 times as many for the 'a'-segment and 5 times as many for the 'å'-segment. The method of cubic-spline interpolation was employed to generate the 45-landmark, F-patterns that are displayed in Fig. 2 and subsequently examined.

4 F-pattern analysis

4.1 Inter-token consistency

It is known that F-patterns exhibit some variability because of the measurement method used, and of one's inability to replicate sounds in exactly the same way. Consequently, the spread magnitude about a token-averaged F-pattern should be useful for gauging measurement consistency, and intrinsic variability to some degree. Table 1 lists spread values that mostly lie within difference-limens for human perception, and are therefore deemed to be tolerable. The spread in F3 for the imitator's "hallå" is relatively large, especially by comparison with his other formants. However, the top left-hand panel of Fig. 2 does show that there is simply greater variability in the F3 of his initial 'a'-segment. Overall, there appear to be no gross measurement errors that prevent a deeper examination of our F-patterns.

Table 1. Inter-token spreads (=standard deviations in Hz) averaged across all 45 landmarks.

	F1	*F2*	*F3*	*F4*
IMITATOR (**SELF**)	33	68	136	72
STOCKHOLM (**STK**)	42	68	28	79
GOTHENBURG (**GTB**)	23	55	71	75
SKANIA (**SKN**)	34	58	36	50
Mean (spread) with IMITATOR:	32 (8)	62 (7)	68 (49)	69 (13)
Mean (spread) without IMITATOR:	33 (10)	60 (7)	45 (23)	68 (16)

4.2 Overview of F-pattern behaviours

For both the imitator's "hallå" and his imitations, there is less curvilinearity in the formant trajectories for the 'a'- and 'l'-segments than in those for the final 'å'-segment, which behaves consistently like a diphthong. The concavity of the F2-trajectory for the Skanian-like 'å'-segment seems to set this dialect apart from the other dialects. Quite noticeably for the 'a'- and 'l'-segments, F1- and F2-trajectories are relatively flatter, and numerically closer to one another than the higher formants. Interestingly again, the F-patterns for the Gothenburg-like "hallå" seem to be more aligned with those corresponding to the imitator's own "hallå".

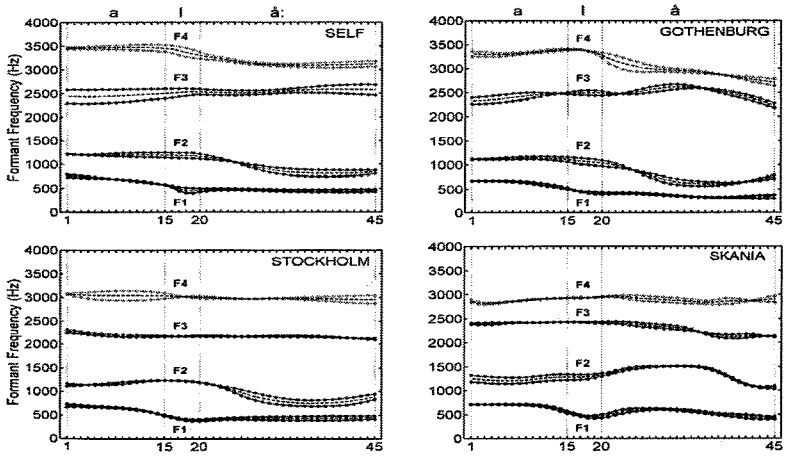

Figure 2. Landmark-normalised F-patterns: Imitator & his imitations of 3 Swedish dialects.

4.3 Imitator versus imitations – a quantitative comparison

The 'a'- and 'l'-segments examined above seem to retain the strongest signature of the imitator's F1- and F2-patterns. To obtain a quantitative verification of this behaviour, we calculated landmark-by-landmark spreads (Fig. 3) of the F-patterns with all data pooled together (left panel), and without the Skania-like data (right panel). The left-panel data highlight a large increase of the spread in F1 and F2 for the final 'å'-segment, thus confirming a major contrast with the other dialectal imitations. The persistently smaller spread in F1 and F2 for the two initial segments raises the hope of being able to detect some invariance in professional imitations of "hallå". The relatively larger spreads in F3 and F4 cast some doubt on these formants' potency for de-coupling our imitator's "hallå" from his imitations.

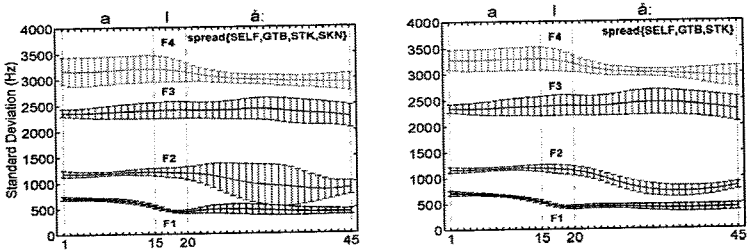

Figure 3. Landmark-by-landmark spreads: (left) all data pooled; (right) Skania-like excluded.

5 Summary and ways ahead

The results of this study are *prima facie* encouraging, at least for the imitations obtained from our professional imitator. It is not yet known whether the near-constancy observed through F1 and F2 of the initial segments of "hallå" will be manifest in other situational tokens, and whether a similar behaviour should be expected with different imitators and phonetic contexts. We have looked at formant-frequencies one at a time but, as shown by Clermont (2004) for Australian English "hello", there are deeper insights to be gained by re-examining these frequencies systemically. The ways ahead will involve exploring all these possibilities.

Acknowledgements

We express our appreciation to Prof. G. Bruce for his auditory evaluation of the imitations. We thank Prof. Bruce and Dr D.J. Broad for their support, and the imitator for his efforts.

References

Clermont, F., 1992. Formant-contour parameterisation of vocalic sounds by temporally-constrained spectral matching. *Proc. 4th Australian Int. Conf. Speech Sci. & Tech.*, 48-53.

Clermont, F., 2004. Inter-speaker scaling of poly-segmental ensembles. *Proc. 10th Australian Int. Conf. Speech Sci. & Tech.*, 522-527.

Rose, P. & S. Duncan, 1995. Naïve auditory identification and discrimination of similar sounding voices by familiar listeners. *Forensic Linguistics 2,* 1-17.

Zetterholm, E., 2003. *Voice imitation: A phonetic study of perceptual illusions and acoustic successes*. Dissertation, Lund University.

Zetterholm, E., D. Elenius & M. Blomberg, 2005. A comparison between human perception and a speaker verification system score of a voice imitation. *Proc. 10th Australian Int. Conf. Speech Sci. & Tech.*, 393-397.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 29–32

Describing Swedish-accented English

Una Cunningham
Department of Arts and Languages, Högskolan Dalarna
uca@du.se

Abstract

This paper is a presentation of the project Swedish accents of English which is in its initial stages. The project attempts to make a phonetic and phonological description of some varieties of Swedish English, or English spoken in Sweden, depending on the status attributed to English in Sweden. Here I show some curious results from a study of acoustic correlates of vowel quality in the English and Swedish of young L1 Swedish speakers.

1 Introduction

1.1 Background

The aim of the proposed project is to document the phonetic features of an emerging variety of English, i.e. the English spoken by young L1 speakers of Swedish. At a time when the relative positions of Swedish and English in Sweden are the stuff of Government bills (Regeringen, 2005), the developing awareness of the role English has as an international language in Sweden is leading to a rejection of native speaker targets for Swedish speakers of English. Throughout what Kachru (1992) called the expanding circle, learners of English are no longer primarily preparing for communication with native speakers of English but with other non-native speakers. In a recent article, Seidlhofer (2005) called for the systematic study of the features of English as a lingua franca (ELF), that is communication that does not involve any native speakers, in order to free ELF from native-speaker norms imposed upon it. She would prefer to see ELF alongside native speaker varieties rather than constantly being monitored and compared to them. The point is that there are features of the pronunciation of native speaker varieties which impede communication, and features of non-native pronunciation which do not disturb communication, and rather than teaching learners to be as native-like as possible, communication would be optimised by instead concentrating on the non-native listener rather than the native listener.

Some young people are British-oriented in their pronunciation, either from RP/BBC English or another accent, others have general American as a clear influence, while another group is not clearly influenced by any native speaker norm. A full phonetic description of these accents of English does not as yet exist, and is of interest as a documentation of an emerging variety of English, at a time when previously upheld targets for the pronunciation of English by Swedish learners have been abandoned and English is growing in importance (Phillipson, 1992; Skolverket, 2006).

1.2 Previous studies

The distinction between English as a Foreign Language (EFL) and English as an International Language (EIL) or English as a Lingua Franca (ELF) is important here. The number of non-native speakers of English increasingly exceeds the number of native speakers, and the native speaker norm as the "given and standard measure" (Jenkins, 2004) for English learners must be questioned. There is a clear distinction between those learners who aspire to sound as