5 Discussion and future work

The synthetic words obtained a reasonable resemblance with the natural words in most cases, and the similarity in age was improved in the second evaluation. The interpolated versions were often judged as older than the intended age in the first evaluation, but in the second evaluation they had become more similar in age to the natural and synthetic versions, indicating that speaker age may be synthesised using data-driven formant synthesis. Still, some of the age estimations were quite unexpected. For instance, the 39, 66 and 69 year olds were judged as much younger than their CA. This may be explained by that these voices were atypical for their age.

One very important point in this study is that synthesis of age by linear interpolation is indeed a crude simplification of the human aging process, which is far from linear. Moreover, while some parameters may change considerably during a certain period of aging (i.e. F_0 and formant frequencies during puberty), others remain constant. Better interpolation techniques will have to be tested. One should also bear in mind that the system is likely to interpolate not only between two ages, but also between a number of individual characteristics, even when the speakers are closely related.

Future work involves (1) improved parameter extraction for formants, (2) better interpolation algorithms, and (3) expansion of the system to handle more speakers (of both sexes), as well as a larger and more varied speech material. Further research with a larger material is needed to identify and rank the most important age-related parameters. If further developed, the prototype system may well be used in future studies for analysis, modelling and synthesis of speaker age and other speaker-specific qualities, including dialect and attitude. The phonetic knowledge gained from such experiments may then be used in future speech synthesis applications to generate more natural-sounding synthetic speech.

References

Boersma, P. & D. Weenink, 2005. *Praat: doing phonetics by computer* (version 4.3.04) [computer program]. Retrieved March 8, 2005, from http://www.praat.org/.

Carlson, R., B. Granström & I. Karlsson, 1991. Experiments with voice modelling in speech synthesis. *Speech Communication 10*, 481–489.

Carlson, R., T. Sigvardson & A. Sjölander, 2002. Data-driven formant synthesis. *Proceedings of Fonetik 2002, TMH-QPSR*, 121–124.

Fant, G., J. Liljencrants & Q. Lin, 1985. A four-parameter model of glottal flow. *STL-QPSR 4*, 1–13.

Hollien, H., 1987. Old voices: What do we really know about them? *Journal of Voice 1*, 2–13.

Jacques, R. & M. Rastatter, 1990. Recognition of speaker age from selected acoustic features as perceived by normal young and older listeners. *Folia Phoniatrica (Basel) 42*, 118–124.

Liljencrants, J., 1968. The OVE III speech synthesizer. *IEEE Trans AU-16*(1), 137–140.

Linville, S.E., 2001. *Vocal Aging*. San Diego: Singular Thomson Learning.

Narayanan, S. & A. Alwan (eds.), 2004. *Text to Speech Synthesis: New Paradigms and Advances*. Prentice Hall PTR, IMSC Press Multimedia Series.

Öhlin, D. & R. Carlson, 2004. Data-driven formant synthesis. *Proceedings of Fonetik 2004*, Dept. of Linguistics, Stockholm University, 160–163.

Xue, S.A. & D. Deliyski, 2001. Effects of aging on selected acoustic voice parameters: Preliminary normative data and educational implications. *Educational Gerontology 21*, 159-168.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 109–112

How do we Speak to Foreigners? – Phonetic Analyses of Speech Communication between L1 and L2 Speakers of Norwegian

Rein Ove Sikveland
Department of Language and Communication Studies,
The Norwegian University of Science and Technology (NTNU), Trondheim
rein.ove.sikveland@hf.ntnu.no

Abstract

The major goal of this study was to investigate which phonetic strategies we may actually use when speaking to L2 speakers of our mother tongue (L1). The results showed that speech rate in general was slower and that the vowel formants were closer to target values, in L2 directed speech compared to L1 directed speech in Norwegian. These properties of L2 directed speech correspond to previous findings for clear speech (e.g. Picheny et al., 1986; Krause & Braida, 2004). The results also suggest that level of experience may influence L2 directed speech; teachers of Norwegian as a second language slowed down the speech rate more than the non-teachers did, in L2 directed speech compared to L1 directed speech.

1 Introduction

When speaking to foreigners in our mother tongue (L1) it might be natural to speak clearer than normal to make ourselves understood, which implies the use of certain phonetic strategies when speaking to these second language learners (L2 speakers).

Previous findings by Picheny et al. (1986) and Krause & Braida (2004) have shown that clear speech can be characterized by a decrease in speech rate, more pauses, relatively more energy in the frequency region of 1-3 kHz, less phonological reductions (e.g. less burst eliminations), vowel formants closer to target values, longer VOT and a greater F0 span, compared to conversational speech. What characterizes L2 directed speech has not been subject to any previous investigations, but one might assume that strategies in L2 directed speech correspond to the findings for clear speech. This has been investigated in the present studies, and the results for speech rate and vowel formants will be presented.

2 Method

To be able to compare speech in L1 and L2 contexts directly, the experiment was carried out by recording native speakers of Norwegian 1) in dialogue with L2 speakers, and 2) in dialogue with other L1 speakers. The dialogue setting was based on a keyword manuscript, to facilitate natural speech, and at the same time be able to compare phonetic parameters in identical words and phonological contexts.

2.1 Subjects

Six native speakers of Norwegian (with eastern Norwegian dialect background) participated as informants. Three of them were teachers in Norwegian as a second language, called P informants (P for "professional"), and three of them were non-teachers, called NP informants (NP for "non-professional"). Six other L1 speakers and six L2 speakers of Norwegian participated as opponents to match each informant in the L1 and L2 contexts. Thus there were 18 subjects participating in the experiment, distributed across twelve recordings.

2.2 Procedure

Recordings were made by placing each informant in a studio while the dialogue opponents were placed in the control room. They communicated through microphones and headphones. The dialogue setting, but not the sound quality, was to represent a phone conversation between two former roommates/partners, and the role of the informants was to suggest to the opponent how to distribute their former possessions, written on a list in the manuscript. There were no lines written in the manuscript, only suggestions of how questions might be asked. The participants were told to carry out the dialogue naturally, but they were not told to speak in any specific manner (e.g. "clearly" or "conversationally"). The speech analyses of the recordings were made using outputs of spectrograms, spectra and waveforms in software "Praat". Only words from the list of possessions were used for analyses, and the corresponding words/syllables/phonemes were measured for each informant in L1 and L2 contexts.

3 Results

3.1 Speech rate

Speech rate was investigated by measuring syllable duration and number of phonemes per second, in ten words for each informant in L1 and L2 contexts (altogether 120 words). The measured words contained four syllables or more. The results showed that syllable duration was longer, and that the number of phonemes per second was lower, in L2 context compared to L1 context. Pooled across informants, the average duration of syllables is 221 ms in L1 context and 239 ms in L2 context. This difference is highly significant ($t (298) = - 4.790$; $p < 0.0001$), and gives a strong general impression that the speech rate is slower in L2 context compared to L1 context.

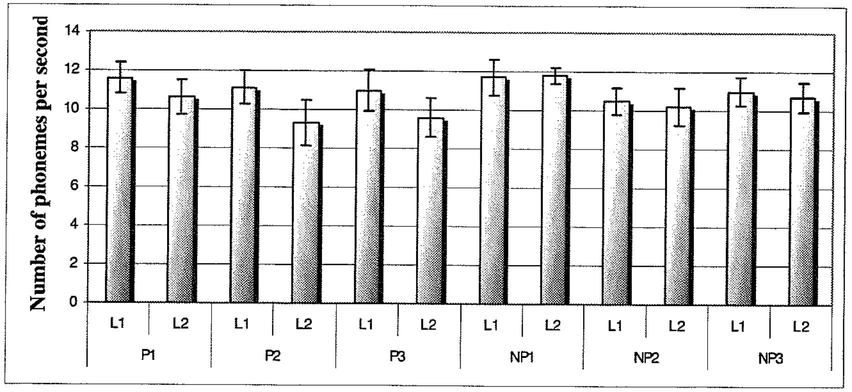

Figure 1. Average number of phonemes per second, for all six informants in L1 and L2 contexts. Error bars are standard deviations.

With the purpose of investigating and describing speech rate more directly, the number of phonemes per second was found to be significantly lower in L2 context compared to L1 context, when pooled across informants ($t (59) = 3.303$; $p < 0.002$). There was an average difference of 0.7 phonemes per second between contexts. In Figure 1 above the number of phonemes per second is shown for all informants in L1 and L2 context. Figure 1 may also show that the speech rate effect is larger for "professional" (P) informants than for "non-professional" (NP) informants. The interaction between level of experience and L1/L2 context on speech rate is significant ($F (1, 58) = 7.337$; $p < 0.009$). Considering these results one has reasons to suggest that speech rate is slower in L2 context compared to L1 context, and that the effect of context on speech rate is dependent on level of experience of the speaker.

3.2 Vowel formants

Formants F1, F2 and F3, in addition to F0, were measured for long and short vowels /a/, /i/ and /u/, representing the three most peripheral vowels in articulation. Since male and female speakers have vocal tracts of different sizes and shapes, the results in Table 1 below are presented for both genders separately. Bold type represents significant differences between contexts, and the results suggest that F1 in /a:/ is generally higher in L2 directed speech than in L1 directed speech, for female ($t (25) = - 3.686$; $p < 0.001$) and male ($t (24) = - 3.806$; $p < 0.001$) speakers. F1 is also significantly higher in L2 context than in L1 context in /a/, for male speakers ($t (20) = - 4.668$; $p < 0.0001$), and in /i/ ($t (19) = - 2.113$; $p < 0.048$) and /u:/ ($t (35) = - 2.831$; $p < 0.008$) for female speakers. A difference in F2 between contexts seems to be evident only for the /i/ vowels, significantly so for male speakers, in /i:/ ($t (23) = - 3.079$; $p < 0.005$) and /i/ ($t (23) = - 5.520$; $p < 0.0001$). F3 values are quite variable within vowels and informants, but significantly higher values in L2 context than in L1 context were found in /i/ ($t (23) = - 2.152$; $p < 0.042$) and /u:/ ($t (35) = - 3.313$; $p < 0.004$) for male speakers.

Table 1. Average values for F1, F2 and F3 in Hz for female (F) and male (M) informants in short and long /a/, /i/ and /u/ vowels. Standard deviations are in parentheses. Bold typing represents statistical significance of differences between L1 and L2 context.

	F1		*F2*		*F3*	
	L1	*L2*	*L1*	*L2*	*L1*	*L2*
/a:/ F (n=26)	**663** (97)	**719** (60)	1165 (89)	1192 (107)	2751 (224)	2724 (165)
M (n=25)	**578** (66)	**632** (46)	1014 (106)	1050 (93)	2562 (209)	2652 (252)
/a/ F (n=20)	729 (121)	738 (67)	1257 (168)	1273 (139)	2728 (226)	2685 (174)
M (n=21)	**552** (77)	**626** (59)	1076 (89)	1088 (123)	2408 (282)	2475 (288)
/i:/ F (n=24)	403 (77)	391 (76)	2362 (269)	2422 (194)	3044 (311)	3035 (299)
M (n=24)	317 (41)	326 (45)	**2029** (124)	**2093** (120)	2947 (261)	3026 (269)
/i/ F (n=20)	**391** (56)	**418** (51)	2287 (229)	2291 (197)	2933 (191)	2951 (154)
M (n=24)	361 (36)	362 (39)	**1933** (109)	**2036** (121)	**2722** (155)	**2795** (229)
/u:/ F (n=36)	**379** (44)	**404** (54)	861 (191)	854 (136)	2728 (209)	2790 (262)
M (n=36)	351 (32)	354 (38)	738 (135)	735 (138)	**2476** (212)	**2572** (199)
/u/ F (n=18)	394 (56)	418 (58)	1028 (164)	1012 (163)	2690 (216)	2638 (225)
M (n=19)	370 (40)	371 (35)	882 (143)	861 (136)	2387 (188)	2388 (146)

If F1 values correlate positively with degree of opening in vowel articulation, the general rise in F1, especially for the /a/ vowels, might be interpreted as a result of a more open mouth/jaw position in L2 context than in L1 context. As suggested by Ferguson & Kewley-Port (2002), a rise in F1 might also be a result of increased vocal effort, which might give an additional explanation to the higher F1 values for /i/ and /u:/. Letting F2 represent the front-back dimension of the vocal tract (high F2 values for front vowels), one might suggest that /i/ vowels (mostly for male speakers) are produced further front in the mouth in L2 context than in L1 context. The tendencies toward higher F2 and F3 frequencies in L2 context compared to L1 context, might indicate that the informants do not use more lip rounding when producing /u/ vowels in L2 context. Rather this point might support our suggestion above that informants in general use a more open mouth position in L2 context than in L1 context.

According to Syrdal & Gopal (1986) one might expect that the relative differences F3-F2 and F1-F0 to describe the front-back and open-closed dimensions (respectively) more precisely than the absolute formant values. In the present investigations, F1-F0 relations led to the same interpretations as for F1 alone, regarding degrees of mouth opening. The F3-F2 relation gave additional information about the vowel /u:/, in that the F3-F2 difference was significantly larger in L2 context than in L1 context (t (40) = - 2.302; p < 0.024). This might be interpreted as /u:/ being produced more back in mouth in L2 context than in L1 context.

Effects of level of experience on formant values or formant relations were not found, which indicates that the differences in vowel formants between L1 and L2 contexts are general among speakers.

4 Conclusions

The results show that L1 speakers modify their pronunciation when speaking to L2 speakers compared to when speaking to other L1 speakers. We have seen that this was so for speech rate, in that the informants had longer syllable durations and fewer phonemes per second in L2 context than in L1 context. The formant values and formant relations indicated that articulation of the peripheral vowels /a/, /i/ and /u/ was closer to target in L2 context compared to L1 context, in both degree of opening and front-back dimensions.

The results for L2 directed speech correspond to those found for clear speech (e.g. Picheny et al., 1986; Krause & Braida, 2004; Bond & Moore, 1994).

Level of experience seemed to play a role in speech rate, in that "professional" L1-L2 speakers differentiated more between L1 and L2 context than "non-professional" L1-L2 speakers did.

References

Bond, Z.S. & T.J. Moore, 1994. A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication 14*, 325-337.

Ferguson, S.H. & D. Kewley-Port, 2002. Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am. 112*, 259-271.

Krause, J.C. & L.D. Braida, 2004. Acoustic properties of naturally produced clear speech at normal speaking rates. *J. Acoust. Soc. Am. 115*, 362-378.

Picheny, M.A., N.I. Durlach & L.D. Braida, 1986. Speaking clearly for the hard of hearing 2: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research 29*, 434-446.

Syrdal, A.K. & H.S. Gopal, 1986. A perceptual model of vowel recognition based on the auditory representation of American English vowels. *J. Acoust. Soc. Am. 79*, 1066-1100.

Working Papers 52 (2006), 113–116

A Switch of Dialect as Disguise

Maria Sjöström[1], Erik J. Eriksson[1], Elisabeth Zetterholm[2], and Kirk P. H. Sullivan[1]

[1] Department of Philosophy and Linguistics, Umeå University
kv00msm@cs.umu.se, erik.eriksson@ling.umu.se, kirk.sullivan@ling.umu.se
[2] Dept. of Linguistics and Phonetics, Centre for Languages and Literature, Lund University
elisabeth.zetterholm@ling.lu.se

Abstract

Criminals may purposely try to hide their identity by using a voice disguise such as imitating another dialect. This paper empirically investigates the power of dialect as an attribute that listeners use when identifying voices and how a switch of dialect affects voice identification. In order to delimit the magnitude of the perceptual significance of dialect and the possible impact of dialect imitation, a native bidialectal speaker was the target speaker in a set of four voice line-up experiments, two of which involved a dialect switch. Regardless of which dialect the bidialectal speaker spoke he was readily recognized. When the familiarization and target voices were of different dialects, it was found that the bidialectal speaker was significantly less well recognized. Dialect is thus a key feature for speaker identification that overrides many other features of the voice. Whether imitated dialect can be used for voice disguise to the same degree as native dialect switching demands further research.

1 Introduction

In the process of recognizing a voice, humans attend to particular features of the individual's speech being heard. Some of the identifiable features that we listen to when recognizing a voice have been listed by, among others, Gibbons (2003) and Hollien (2002). The listed features include *fundamental frequency (f0), articulation, voice quality, prosody, vocal intensity, dialect/sociolect, speech impediments and idiosyncratic pronunciation.* The listener may use all, more, or only a few, of these features when trying to identify a person, depending on what information is available. Which of these features serve as the most important ones when recognizing a voice is unclear. Of note, however, is that, according to Hollien (2002), one of the first things forensic practitioners look at when trying to establish the speaker's identity is dialect.

During a crime, however, criminals may purposely try to hide their identity by disguising their voices. Künzel (2000) reported that the statistics from the German Federal Police Office show that annually 15-25% of the cases involving speaker identification include at least one type of voice disguise: some of the perpetrators' 'favourites' include: falsetto, pertinent creaky voice, whispering, faking a foreign accent and pinching one's noise. Markham (1999) investigated another possible method of voice disguise, *dialect imitation.* He had native Swedish speakers attempt to produce readings in various Swedish dialects that were not their native dialects. Both the speaker's ability to consistently keep a natural impression and to mask his or her native dialect were investigated. Markham found that some speakers are able to successfully mimic a dialect and hide their own identity. Markham also pointed out that to