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.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
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 idiosynctratic 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

avoid suspicion it is as important to create an impression of naturalness, as it is to hide one's identity when using voice disguise.

In order to baseline and delimit the potential impact on speaker identification by voice alone due to dialect imitation a suite of experiments were constructed that used a native bidialectal speaker as the speaker to be identified. The use of a native bidialectal speaker facilitates natural and dialect consistent stimuli. The four perception tests presented here are excerpted from Sjöström (2005). The baselining of the potential problem is of central importance for forensic phonetics since, if listeners can be easily fooled, it undermines earwitness identification of dialect and suggests that forensic practitioners who currently use dialect as a primary feature during analysis would need to reduce their reliance on this feature.

2 Method

Four perception tests were constructed. The first two tests investigated whether the bidialectal speaker was equally recognizable in both his dialects. The second two tests addressed whether listeners were distracted by a dialect shift between familiarization and the recognition task.

2.1 Speech material

The target bidialectal speaker is a male Swede who reports that he speaks Scanian and a variety of Stockholm dialect on a daily basis. He was born near Stockholm but moved to Scania as a five-year old. An acoustic analysis of the speaker's dialect voices was performed, which confirmed that his two varieties of Swedish carry the typical characteristics of the two dialects and that he is consistent in his use of them.

Two recordings of *The Princess and the Pea* were made by the bidialectal speaker. In one of them he read the story using the Stockholm dialect, and in the other he read it using his Scanian dialect.

Four more recordings of *The Princess and the Pea* were made; two by two male mono-dialectal speakers of the Stockholm dialect (ST) and two by two male mono-dialectal speakers of the Scanian dialect (SC). These speakers (hereafter referred to as foils) were chosen with regard to their similarities with the target voice in dialect, age, and other voice features such as creakiness. For further details, see Sjöström (2005).

2.2 The identification tests

Four different earwitness identification tests were constructed for participants to listen to. Each test began with the entire recording of *The Princess and the Pea* as the familiarization voice, and was followed by a voice line-up of 45 stimuli. The 45 stimuli consisted of three phrases selected from each recording presented three times for each *speaker* (3 x 3 x 5 = 45). Each voice line-up contained the four foil voices and one of the target's two dialect voices (see Table 1). For example, the test 'SC-ST' uses the target's Scanian voice as the familiarization voice and the target's Stockholm dialect voice in the line-up. Test SC-SC and Test ST-ST were created as control tests. They afford investigation of whether the target's Stockholm and Scanian dialects can be recognized among the voices of the line-up, and to test if the two different dialects are recognized to the same degree. Tests ST-SC and SC-ST investigate if the target can be recognized even when a dialect shift between familiarization and recognition occurs.

80 participants, ten in each listener test, took part in this study. All were native speakers of Swedish and reported no known hearing impairment. Most of the listeners were students at either Lund University or Umeå University, and all spoke a dialect from the southern or northern part of Sweden.

Table 1. The composition of the voice identification tests showing which of the target's voices was used as familiarization voice and which voices were included in the voice line-up for each of the four tests.

Test	*Familiarization voice*	*Line-up voices*
SC-SC	TargetSC	Foil 1-4 + TargetSC
ST-ST	TargetST	Foil 1-4 + TargetST
ST-SC	TargetST	Foil 1-4 + TargetSC
SC-ST	TargetSC	Foil 1-4 + TargetST

2.3 Data analysis

In this yes-no experimental design responses can be grouped into four different categories: *hit* (when the listener correctly responds 'yes' to the target stimulus), *miss* (when the listener responds 'no' to a target stimulus), *false alarm* (when the listener responds 'yes' to a non-target stimulus) and *correct rejection* (when the listener correctly responds 'no' to a non-target stimulus). By calculating the hit and false alarms rates as proportions of the maximum possible number of hits and false alarms, the listeners' *discrimination sensitivity* can be determined, measured as *d'*. This measure is the difference between the hit rate (H) and the false alarm rate (F), after first being transformed into z-values. The d'-equation is: $d' = z(H)-z(F)$ (see Green & Swets, 1966).

3 Results and discussion

Participants of the control tests, SC-SC and ST-ST, show positive mean *d'*-values (1.87 and 1.93). It was shown through a two-tailed Student's t-test that there was no significant difference in identification of the two dialects and they can therefore be considered equally recognizable ($t(38)=-0.28$, $p>0.05$). By conducting a one-sample *t*-test it was shown that the *d'*-values for both tests are highly distinct from 0 ($t(39)=18.45$, $p<0.001$) and therefore high degree of identification of both dialects can be concluded.

The responses for the dialect shifting tests, ST-SC (mean d' = 0.44); SC-ST (mean d' = -0.07), did not significantly differ ($t(38)=1.93$, $p>0.05$). The target voice in these two tests can be considered equally difficult to identify. A one-sample t-test was conducted and showed that the mean d'-value of the two tests was not significantly separated from 0 ($t(39)=1.36$, $p>0.05$), indicating random response. Combining the responses for the 'control tests' (ST-ST; SC-SC) and the 'dialect shifting tests' (ST-SC; SC-ST) and comparing the results to each other revealed a significant difference between the two test groups ($t(78)=5.97$, $p>0.001$) (see Fig. 1). Thus, dialect shift has a detrimental effect on speaker identification.

4 Conclusions

The results indicate that the attribute dialect is of high importance in the identification process. It is clear that listeners find it much more difficult to identify the target voice when a shift of dialect in the voice takes place. One possible reason for the results is that when making judgments about a person's identity, dialect as an attribute is strong and has a higher priority than other features.

The baselining of the potential problem we have conducted here shows that a switch of dialect can easily fool listeners. This undermines earwitness identification of dialect and suggests that forensic practitioners who currently use dialect as a primary feature during analysis need to reduce their reliance on this feature and be aware that they can easily be misled.

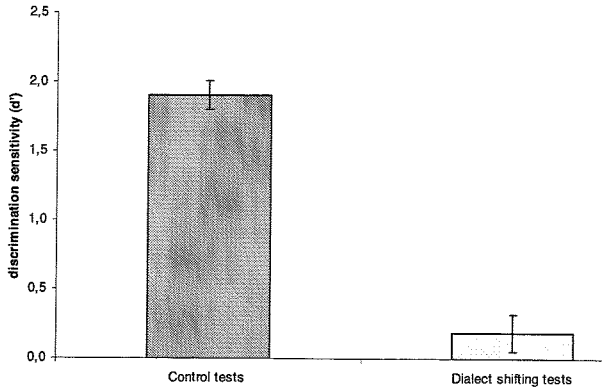

Figure 1. Mean discrimination sensitivity (d') and standard error for Control tests (SC-SC and ST-ST combined) and Dialect shifting tests (ST-SC and SC-ST combined).

If used as a method of voice disguise, a perpetrator could use one native dialect at the time of an offence and use the other in the event of being forced to participate in a voice line-up as a suspect. Needless to say this method of voice disguise could have devastating effects on witness accuracy as they would not able to recognize the perpetrators voice when using different dialect, or yet worse, that the witness would make an incorrect identification and choose another person whose dialect is more similar to the voice heard in the crime setting.

In order to assess whether voice disguise using imitated dialect can have as drastic an impact upon speaker identification as voice disguise by switching between native dialects, research using imitated dialect as a means of disguise is required.

Acknowledgements

Funded by a grant from the Bank of Swedish Tercentenary Foundation Dnr K2002-1121:1-4 to Umeå University for the project 'Imitated voices: A research project with applications for security and the law'.

References

Gibbons, J., 2003. *Forensic Linguistics*. Oxford: Blackwell Publishing.

Green, D.M. & J.A. Swets, 1966. *Signal detection theory and psychophysics*. New York: John Wiley and sons, Inc.

Hollien, H., 2002. *Forensic voice identification*. San Diego: Academig Press.

Künzel, H.J., 2000. Effects of voice disguise on speaking fundamental frequency. *Forensic Linguistics 7*, 1350-1771.

Markham, D., 1999. Listeners and disguises voices: the imitation and perception of dialectal accent. *Forensic Linguistics 6*, 289-299.

Sjöström, M., 2005. *Earwitness identification – Can a switch of dialect fool us?* Masters paper in Cognitive Science. Unpublished. Department of Philosophy and Linguistics, Umeå University.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 117–120

Prosody and Grounding in Dialog

Gabriel Skantze, David House, and Jens Edlund
Department of Speech, Music and Hearing, KTH, Stockholm
{gabriel|davidh|edlund}@speech.kth.se

Abstract

In a previous study we demonstrated that subjects could use prosodic features (primarily peak height and alignment) to make different interpretations of synthesized fragmentary grounding utterances. In the present study we test the hypothesis that subjects also change their behavior accordingly in a human-computer dialog setting. We report on an experiment in which subjects participate in a color-naming task in a Wizard-of-Oz controlled human-computer dialog in Swedish. The results show that two annotators were able to categorize the subjects' responses based on pragmatic meaning. Moreover, the subjects' response times differed significantly, depending on the prosodic features of the grounding fragment spoken by the system.

1 Introduction

Detecting and recovering from errors is an important issue for spoken dialog systems, and a common technique for this is verification. However, verifications are often perceived as tedious and unnatural when they are constructed as full propositions verifying the complete user utterance. In contrast, humans often use fragmentary, elliptical constructions such as in the following example: "Further ahead on the right I see a red building." "Red?" (see e.g. Clark, 1996).

In a previous experiment, the effects of prosodic features on the interpretation of such fragmentary grounding utterances were investigated (Edlund et al., 2005). Using a listener test paradigm, subjects were asked to listen to short dialog fragments in Swedish where the computer replies after a user turn with a one-word verification, and to judge what was actually intended by the computer by choosing between the paraphrases shown in Table 1.

Table 1. Prototype stimuli found in the previous experiment.

Position	Height	Paraphrase	Class
Early	Low	Ok, red	ACCEPT
Mid	High	Do you really mean red?	CLARIFYUNDERSTANDING
Late	High	Did you say red?	CLARIFYPERCEIVE

The results showed that an early, low F_0 peak signals acceptance (display of understanding), that a late, high peak is perceived as a request for clarification of what was said, and that a mid, high peak is perceived as a request for clarification of the meaning of what was said. The results are summarized in Table 1 and demonstrate the relationship between prosodic realization and the three different readings. In the present study, we want to test the hypothesis that users of spoken dialog systems not only perceive the differences in prosody of synthesized fragmentary grounding utterances, and their associated pragmatic meaning, but that they also change their behavior accordingly in a human-computer dialog setting.