Lund University, Dept. of Linguistics
Working Papers 49 (2001), 86–89

Voluntary and involuntary speech variations - a few examples from the VeriVox database

Inger Karlsson
Dept. of Speech, Music and Hearing, KTH
inger@speech.kt.se

Abstract

Some speech variations due to involuntary and voluntary speech productions have been investigated. In this very preliminary report duration variations for two speakers are discussed.

1 Introduction

The aim of the VeriVox project (Karlsson et al. 2000) was to improve the reliability of automatic speaker verification (ASV) by developing novel, phonetically informed methods for coping with the variation in a speaker's voice. To get a speech material to test methods on, special software was developed. The software was designed to systematically elicit different types of voluntary and involuntary speech variation. A database containing speech from 50 Swedish male speakers was collected using this software.

The ultimate aim of the project was to suggest methods to take care of within-speaker variations in automatic speaker verification. To achieve this the voluntary variations that are elicited during enrolment should cover the variations that may occur in normal situations. It has been presumed in this and previously published papers (Karlsson et al 1998, Karlsson 1999) that the elicited involuntary speech variations actually cover some of these normal variations.

It was shown in the ASV experiments performed within VeriVox (Karlsson et al 2000) that including the voluntary speech variations in the enrolment set enhanced the performance on the involuntary speech variations. One question that remains to be answered is if all voluntary variations are of help. To shed some light one this, an investigation of the acoustic variations have been performed on voluntary and involuntary speech variations.

2 Speech elicitation

The speech database was recorded using a prototype version of eliciting software developed within the project by the partner in Geneva. The software is designed to systematically elicit different types of voluntary and involuntary speech variation. The users are sitting in front of a computer screen on which instructions for the different tasks appeared. Explanations of the different parts of the recording sessions are also given on the screen. Voluntary speech variation is elicited by directing the user to deliberately speak in a number of different modes, including Normal, Fast, Slow, Weak, Strong and Denasalised speech (pinched nose).

The software elicits involuntary variation by means of an interactive module in which users perform a succession of tasks, which cause them to speak normally, faster, and louder without being explicitly asked to do so. The tasks include (i) speaking in the presence of two levels of background white noise (administered through headphones), (ii) speaking from memory at an increased rate due to time pressure and (iii) speaking while solving a divided attention logical reasoning and auditory recognition task, with background noise distraction, allowing the recording of stressed speech. Non-directed normal speech samples are also collected as part of this interactive module. All these tasks are designed to elicit the types of involuntary speech variation, which might realistically occur in use of speaker verification systems. This second module (involuntary variation) of the elicitation system uses the same digit sequences and phrases as used in the first part (voluntary variation). The users are also asked to indicate their stress level on a scale from 0 to 9 after each task.

2.1 Speakers

The eliciting software was used for collecting a database with 50 male Swedish speakers. For each speaker a single 30-minute session, which includes both enrolment and verification utterances for the speaker was recorded. Given that our interest is mainly in speaker variations due to systematical changes in factors like speaking rate, loudness level and formant frequencies, it was found reasonable to use material from a single recording session in a pilot study. Material from only two of these 50 speakers will be discussed in this paper. Further discussions of the speech database can be found in earlier publications (Karlsson et al. 1998, Karlsson et al. 2000).

3 Measurements

The segment durations are measured automatically using in-house alignment software (Sjölander 2001) and the segment boundaries are then corrected by hand. Data for all 50 speakers will be available. So far segment durations for all utterances by two speakers have been studied. In this paper data for the phrase [_d__ta ær _sl___t_t po_ ___p___ft <number sequence>] that occurs 14 times in the speech material and a six-digit string are discussed. This string occurred both as an isolated ID sequence and in repetitions of digit strings. The studied material occurred in the following contexts:

(1) Voluntary: normal, weak voice, loud voice, slow speech, fast speech, pinched nose (= de-nasalised) speech.

(2) Involuntary: ID number: the very start of the session, in loud noise, after having repeated street names from memory under time constraints, after high cognitive load task, that is split attention due to logical reasoning and auditory recognition task. Phrase: after memorising street names, in weak noise, in loud noise, after repeating street names from memory, after reading without voicing, after repeating street names from memory under time constraints, after high cognitive load task

The two speakers were chosen because they showed a larger than average duration variation in earlier studies. Speaker A indicated a low degree of stress, (4-5), while Speaker B was more stressed, (5-8), by the different tasks.

Results from the measurements for the phrase are given in Figure 1. For each phoneme a mean duration was calculated. The quotients between the individual durations and the mean were calculated to get a relative duration. The data points in the figure were then calculated as the mean of the relative durations for all vowel and consonant segments in the phrase context indicated in Figure 1.

4 Discussion

When comparing the voluntary and involuntary productions for the two speakers processed so far it seems that the slow speech is not reproduced in any of the involuntary variations. On the other hand the voluntary fast speech durations coincided with some of the involuntary variations. This can be seen for both speakers in Figure 1 and the same was found for the ID digit sequence. For both speakers the vowels tend to vary more in length than the consonants. In Figure 1 both speakers show longer than normal segments in the voluntary loud utterances which agrees with what has been found in involuntary loud speech due to background noise (Junqua 1995). The utterances in both weak and loud noise though the segments are shorter than the mean. This may depend on that the data come from the part of the phrase before the digit sequence that is the stressed part of the sentence. The data for the ID digit sequence show longer segment durations for the speech in noise condition.

The results are so far very preliminary, more speakers are being processed. More data will be presented at the meeting.

Acknowledgement

The VeriVox project was supported by the EU Esprit program. Coordinator of the project was KTH (S), and partners were Cambridge Univ. (UK), CNRS (F), Queen Margaret College (UK), Univ. Of Dublin (IRL), Univ. Bonn (D), Univ. Geneva (CH), and Ensigma Ltd. (UK).

References

Junqua, J.-C. 1995. ‘The influence of acoustics on speech production: a noise-induced stress phenomenon known as the Lombard reflex.’ *Proc. of the ESCA-NATO Tutorial workshop on Speech under Stress*, 83-90

Karlsson I, Banziger T, Dankovicová J, Johnstone T, Lindberg J, Melin H, Nolan F & Scherer K. 1998. ‘Within-speaker variability due to speaking manners.’ In: Mannell RH & Robert-Ribes J, eds. *Proc of ICSLP98, Intl Conference on Spoken Language Processing*, Sydney, Australia, 1998, 2379-2382.

Karlsson I., Banziger T., Dankovicova J., Johnstone T., Lindberg J., Melin H., Nolan F., Scherer K. 2000. ‘Speaker Verification with Elicited Speaking-styles in the VeriVox project’, *Speech Communication 31*, 121-129.

Karlsson I. 1999. ‘Within-speaker variability in the VeriVox database.’ In: Andersson R, Abelin Å, Allwood J & Lindblad P, eds. *Proc of Fonetik 99, 12th Swedish Phonetics Conference,* June 2-4, 1999, Göteborg, Sweden, 93-96.

Sjölander, K. 2001. ‘Automatic alignment of phonetic segments.’ In this volume

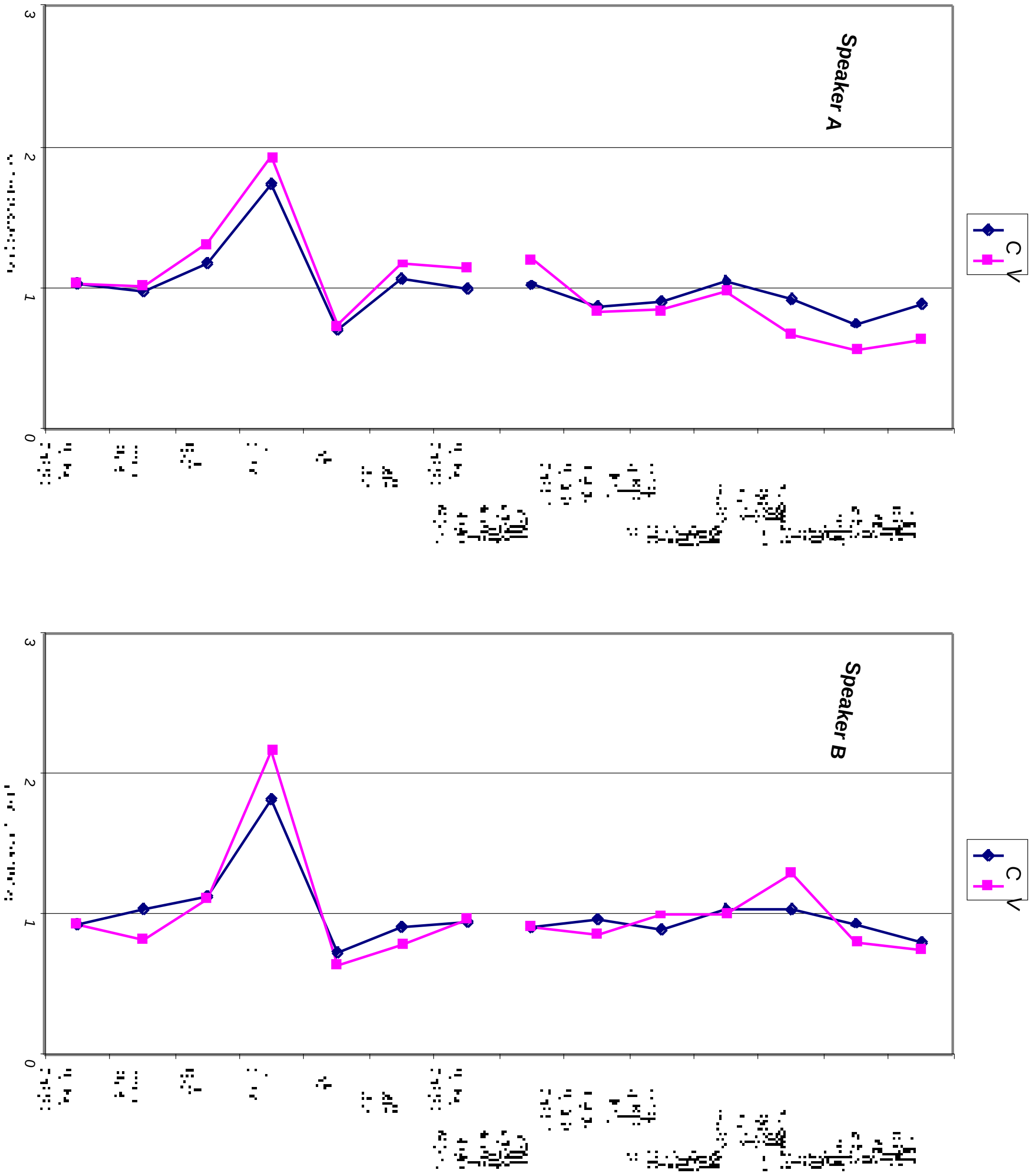

Figure 1. Variations in segment duration in different voluntary, right part, and involuntary, left part, speech productions for two speakers. The data pertains to the phrase start [_d__ta ær _sl___t_t po_ ___p___ft] that is *This is the end of task...*