Within-speaker variability due to speaking manners

I. Karlsson[1], T. Banziger[3], J. Dankovicová[2], T. Johnstone[3], J. Lindberg[1], H. Melin[1], F. Nolan[2], K. Scherer[3]

(1) KTH Department of Speech, Music and Hearing, Stockholm, Sweden
(2) CULD Department of Linguistics, University of Cambridge, Cambridge, UK
(3) FAPSE Department of Psychology, University of Geneva, Geneva, Switzerland

ABSTRACT

Some preliminary investigations of within-speaker variations due to voluntary and induced speaking manners have been performed. The ultimate aim of the investigations was to suggest methods to take care of within-speaker variations in automatic speaker verification. Special software was developed to systematically elicit different types of voluntary and involuntary speech variations that might realistically occur in every-day situations. A database containing speech from 50 Swedish male speakers was collected using this software. Acoustic analyses have been performed on and the results compared between voluntary and involuntary speech variations. The acoustic parameters that have been studied included segment durations, formant frequencies at vowel midpoints, fundamental frequency and overall amplitude and amplitude in frequency bands.

1. INTRODUCTION

The present paper presents some methods and data attained in the ESPRIT-VeriVox project. The ultimate aim of the VeriVox project 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. Up to now, ASV research has treated within-speaker variation as if it were random, but in fact phonetic research reveals that within-speaker variation is highly structured. Different speaking rates, loudness levels, styles, emotional states, and so on, all cause predictable changes in the acoustic speech signal. The (long-term) goal is to exploit such known phonetic and phonological regularities to reduce the false rejection rate in ASV without a concomitant rise in the risk of false acceptances. This paper reports on some of the results achieved during the six month project. These include software for eliciting different speaker behavior, a database of various elicited speaking-styles and an acoustic analysis. The elicited speaking styles were decided to cover the needs of our current approach to using phonetic knowledge in an ASV system. They were used to form an enrolment set for the ASV system that was called "structured training". This approach has been tested using a state of the art HMM-based ASV system developed in the CAVE project [1]. The ASV experiment is more fully reported in [2].

2. DATABASE

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

The software elicited involuntary variation by means of an interactive module in which users performed a succession of tasks, which caused them to speak normally, faster, and louder without being explicitly asked to do so. The tasks included (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 were 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 used the same digit sequences as used in the first part (voluntary variation). The users were 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. All speakers in the database come from the same (broad) dialect region around Stockholm. The age range was 22 to 78 years with a concentration between 25 and 40. Nearly all speakers had some computer experience. As the task performed to elicit stressed speech was similar to a computer game, they were asked if they were used to play computer games.

2.2. Speech material

The database was constructed to contain speech to enroll speakers into an ASV system with neutral and with structured training, and further to test the system with a variety of speaking-styles. The speech was recorded in a sound treated booth using a headset, Sennheiser HME 25-1, with a condenser microphone and closed headphones. The speech is sampled with a high sampling rate to allow for various acoustic analyses of

the speech. The following speech material was recorded during the different tasks as described below.

1. A sequence of digits consisting of six digits in six different orders and a short phrase were read in different manners: Neutral, Weak, Strong, Slow, Fast, Denasalised (pinched nose), Neutral, always in that order. The six digits were: 2 [tv'o:], 3 [tr'e:], 4 [f"y:ra], 5 [f'ɛm], 7 [ʂ'ɯ:] or [ʂ'ɯ:], 0 [n'ɔl]. The phrase was: " Detta är slutet på uppgift 235047", in English "This is the end of task 235047". The digit sequences were constructed so that every digit appeared in all possible contexts: 024753, 503427, 237054, 304572, 407325, 743520

2. A list of six Stockholm street names that were read and memorized. The streets were well known and ordered from north to south to simplify recall.

3. Speech in noise; the subject was asked to describe a concept while white noise was presented over headphones. After 30 seconds' exposure to noise, the subject was asked to repeat a phrase containing a number sequence and an ID-number. The phrase and the ID-number were recorded. The task was repeated with a higher noise level. The noise level was not measured but was experienced to be *loud* but not *unacceptably loud* by the subjects.

4. The street list was repeated from memory.

5. A silent task, only initial and final phrases with task numbers were recorded.

6. The street list was repeated from memory with a strict time limit of 10 seconds.

7. During a logic reasoning task with two different telephones sounding, the subjects were asked to answer their telephone by reading a digit sequence shown on the screen. The sequence was chosen randomly from the six sequences recorded during the first task. The number sequence was recorded. The reasoning task consisted of deciding if a proposal was true given a set of postulates. For example: Postulates: A: Fishes can fly as they have wings. B: Storks don't have wings. Statements: (i) It is possible to fly with wings (ii) Without wings one can't fly through the air (iii) Storks can't fly (iv) Fishes and storks have nothing in common (v) All fishes are storks

Between tasks the phrases "Detta är början på uppgift <number sequence>" in English: "This is the beginning of task < >" and "Detta är slutet på uppgift <number sequence>", in English: "This is the end of task < >", are recorded, sometimes also the ID-code 503427. The speech signal was recorded on the computer hard disc with 22.1 kHz sampling rate. The acoustic analysis was performed on these high quality recordings while the ASV experiments were performed on a telephone quality version.

3. ACOUSTIC ANALYSIS

3.1 Pilot study of six speakers

Segment durations and formant frequencies at vowel midpoints were measured for 6 speakers saying 3-0-4-5-7-2 spoken in seven conditions: Neutral, Loud, Weak, Slow, Denasal (pinched nose), and (cognitively) Stressed. The six speakers all came from the Stockholm area and their age ranged from 25 to 40 years. One token of each word in each condition was analyzed for each speaker, except in the case of Neutral and Stressed where two tokens were analyzed. For the same six speakers, the durations for another six-digit string uttered in the 'Memory under time pressure' task was compared with the same string uttered in the Neutral condition. Despite the small amount of data a number of trends emerge, some of which are summarized below. As expected, all segments in Slow are longer, and most in Fast are shorter. Loud and Weak predominantly involve longer segments. Stressed and 'Memory under time pressure', seem to involve almost consistent shortening of segments. If each segment's duration is expressed as the percentage change it undergoes as a proportion of the utterance, relative to Neutral, it emerges that a rate change is unevenly distributed over different categories of sound. In Slow there is a clear tendency for the vowels to take up a greater proportion of the lengthening and the consonants less, relative to Neutral; this is also true for Loud. The pattern in Fast is reversed: several consonants take up a greater proportion of the utterance and several vowels take up a smaller proportion.

The first and second formants were also measured. In Fast, with the exception of /fɛm/, the vowels are mid-centralized. In Fast speech there is perhaps less time for the tongue to achieve peripheral articulations. In Slow the vowels are more peripheral. Stressed shows on a smaller scale the (de-peripheralisation) pattern of Fast, while Loud shows in some vowels a pattern of peripheralisation, like Slow.

The styles tested bring about radical restructuring of the temporal and spectral properties of the speech. This gives a clue to why errors arise in the verification process. Given that 'claim' utterances may differ durationally in complex ways, then even if it is possible to 'time-warp' the claim utterances so that they align well with Neutral reference data, the aligned segments (vowels in particular) will match badly in spectral terms.

3.2. Influence of computer experience

The speakers have been divided into two groups according to if they used to play computer games, here called *players*, or not, called *non-players*. This was done as it was suspected that the *players* would be considerably less stressed during the last task. This suspicion was tested with the self-assessed stress-levels that the subjects indicated after each task, see Table 1. The *players* have on average indicated a lower stress level than the *non-players*, and in particular the difference for the last task is large. For the *players*, the recalling of street names under time pressure is nearly as stressful as the logical reasoning task while for the *non-players* the difference is considerable.

Condition	Players	Non-players
After reading in different manners	4.50	4.81
After reading and memorizing street names	4.94	5.29
After talking in noise	5.11	5.33
After recalling street names	5.17	5.81
After answering an inquiry, no speech	4.94	5.38
After recalling street names in 10 sec	5.67	6.05
After logic reasoning	5.89	7.10

Table 1. Average stress levels. The speakers are divided into two groups, *players* who have played computer games, and *non-players* who have no experience of computer games.

For these two groups some examples of the sequence 503427 have been extensively investigated and compared. This particular sequence was chosen as it served as an ID-number for all speakers and occurred frequently in the material prompted in the same way. It also often occurred in the last task where the six sequences were randomized. The samples investigated were the very first utterance that was recorded during the recording session, a recording with loud noise in the earphones, a recording after the memory task and the last utterance recorded (the subjects did not know it was the last.) The sequence sometimes occurred more than once in the logic reasoning task. If so, an average was calculated for that speaker. All utterances have been compared with the sequence uttered after the memory task. That utterance is called Neutral in the following discussion. For various reasons some samples where missing for 9 of the speakers. Of the remaining 41 speakers, 19 belong to the group *players* and 22 *non-players*.

The duration was measured for all segments. The power below and above 1 kHz was measured for all voiced segments and the difference between the two values was calculated. This gives both a demonstration of the overall amplitude variations and some indications of voice source differences. The fundamental frequency was measured at extreme points through the utterances. Some speakers had fairly creaky voices, they were not included. Typical for these speakers was that the utterances spoken against background noise were not creaky.

Duration. The two groups showed different patterns in word duration changes, see Figure 1. For both groups words uttered in noise were considerably longer than the Neutral utterance. The same effect has been found in previous investigations, [3]. The words spoken during the logic reasoning task (the task most similar to computer games) were longer compared to the Neutral utterance for the non-player group while they were shorter for the *players*. In all cases there is a lengthening of the final word. The difference between the two groups also showed in a comparison between the Neutral and the first utterance. For the *non-players* all words except the last in the first utterance were considerably longer (10%) while there were only small differences for the *players*.

The only instruction the speakers were given was to read aloud what was shown on the screen. The number sequences were written on one line and were equally spaced. Still, nearly all speakers grouped the numbers three by three. This resulted in a pause in the middle of the sequence. In the Noise utterance this pause was considerably longer, about 30%, than in the Neutral utterance, while words and segments were about 15% longer. Most of the lengthening in the words occurred in the voiced segments, both vowels and consonants. This complies with the findings for the 6 speakers discussed above. In the Stressed utterances by the player group most of the shortening occurred in the voiced and voiceless consonants. The same occurred in the shortened words in the same utterances by the *non-players* while lengthening occurred in vowels. The pause was shortened by nearly 20% for both groups.

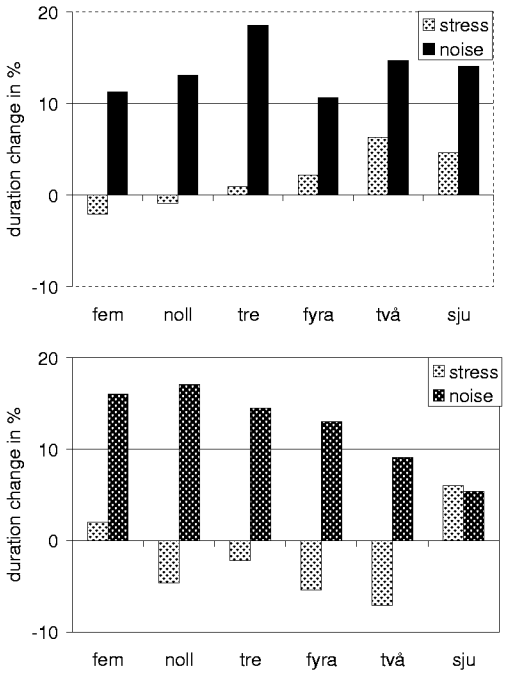

Figure 1. Word duration for the Stressed and Noise utterances in percent of the Neutral utterance. Data for *non-players* are shown above, for *players* below.

Amplitude. The amplitude below and above 1 kHz was measured in the middle of all voiced segments. The amplitude below 1 kHz is for voiced segments about equal to the total amplitude and difference between these two measures gives an estimate of the voice source slope. The recordings were not calibrated so only differences in amplitude have been investigated. The two groups *players* and *non-players* showed very similar values. The Noise utterances were about 4 dB louder than the Neutral utterance. A similar difference was found above 1 kHz, which indicates that the voice source slope is less steep in the Noise utterances. The Stressed utterances were about 1 dB louder than the Neutral utterance. The *non-players* showed a weak tendency towards more energy above 1 kHz for the Stressed utterances while no such tendency was found for the *players*.

Fundamental frequency. Peak values and values in the midpoints of some vowels that did not contain extreme values were measured in all utterances. Only very small differences were found between different type of utterances. The Noise utterance showed about 5% higher values than the Neutral. The Stressed utterance differed slightly between the two groups. The *non-players* showed about 3% higher fundamental frequency through the utterance compared to the Neutral utterance while there was no difference for the *players*.

Formant frequencies. Formant frequencies have been measured at vowel midpoints for a few speakers. The differences between the different utterances does not seem very large, but see the following section for a discussion of some speakers that created problems in the speaker verification experiment.

3.3. ASV problem speakers

The speech database studied in this paper have been used in a speaker verification experiment, [2]. In the experiment 31 tests were made for each speaker. The overall results can be summarized in EER, the equal error rate, where the false acceptance rate is equal to the false rejection rate. In our experiment AVERAGE EER was 2.7%. A few speakers performed considerably worse than that. Two of those speakers have been investigated more closely. One speakers was not accepted as himself 4 times out of 31 trials and 6 other speakers were accepted as him by the system. The other speaker was rejected 3 times out of 31 trials and 2 other speakers were accepted as him. Both these speakers show larger differences between the different utterances in all acoustic parameters that have been studied in this investigation. The larger variations in formant freuqencies are demonstrated in Figure 2 where formant frequencies for different utterances are displayed. Another example is that the fundamental frequency contours for most speakers were very similar between the different conditions, while for the two problem speakers the contours varied considerably.

4. CONCLUSION

The aim of the VeriVox project was to find methods to handle within-speaker variations due to different types of stress in ASV. An important part was to elicit and analyze within-speaker variation. In the present report, some steps towards a description of the within-speaker acoustic variations in the elicited speaking styles have been made. So far common trends have been discussed. The between speaker variations are large though. More work will be performed to describe these.

5. 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).

Special thanks to Lennart Nord (KTH) for his assistance with the database recording.

6. REFERENCES

1. Bimbot F., Hutter H.-P., Jaboulet C., Koolwaaij J., Lindberg J., and Pierrot J.-B. "Speaker Verification in the Telephone Network: Research activities in the CAVE Project," Proc. EUROSPEECH'97, 971-974, 1997

2. Karlsson, T. Banziger, J. Dankovicova, T. Johnstone, J. Lindberg, H. Melin, F. Nolan, K. Scherer, "Speaker Verification with Elicited Speaking-styles in the VeriVox project". Proc. of RLA2C ("Speaker Recognition and its Commercial and Forensic Applications") 207-210, 1998

3. Junqua, J.-C. "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, 1995

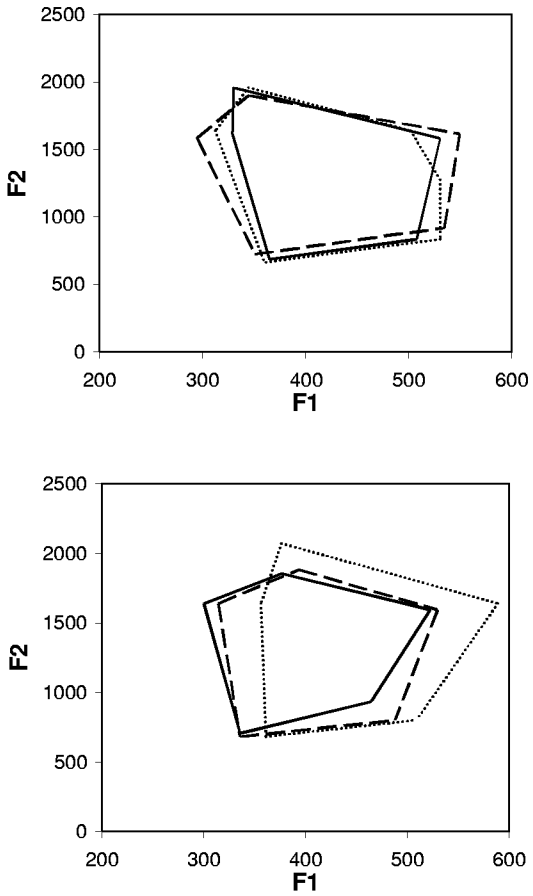

Figure 2. Average formant frequencies for 5 speakers that were recognized by the ASV system, above, and the two problem speakers, below. Three utterances are represented in the Figure, Neutral is signified by a solid line, Stressed by a dashed line and Noise by a dotted line.