CAN AUTOMATIC SPEAKER VERIFICATION BE IMPROVED BY TRAINING THE ALGORITHMS ON EMOTIONAL SPEECH?

K. R. Scherer, T. Johnstone, G. Klasmeyer, & T. Bänziger
University of Geneva, Switzerland

ABSTRACT

The ongoing work described in this contribution attempts to demonstrate the need to train ASV algorithms on emotional speech, in addition to neutral speech, in order to achieve more robust results in real life verification situations. A computerized induction program with 6 different tasks, producing different types of stressful or emotional speaker states, was developed, pretested, and used to record French, German, and English speaking participants. For a subset of these speakers, physiological data were obtained to determine the degree of physiological arousal produced by the emotion inductions and to determine the correlation between physiological responses and voice production as revealed in acoustic parameters. In collaboration with a commercial ASV provider (Ensigma Ltd.), a standard verification procedure was applied to this speech material. This paper reports the first set of preliminary analyses for the subset of 30 German speakers. It is concluded that an evaluation of the promise of training ASV material on emotional speech requires in-depth analyses of the individual differences in vocal reactivity and further exploration of the link between acoustic changes under stress or emotion and verification results.

1. INTRODUCTION

This contribution summarizes the current state of the EMOVOX project at the University of Geneva (funded by the Swiss National Research Fund) and reports first results. The overall aim of the project, which has grown out of the European VERIVOX research consortium [1] is to systematically quantify and improve understanding of the pattern of changes that different physiological and psychological speaker states produce in the acoustic speech signal, leading to the construction of a database on intra-speaker variation in speech with an emphasis on speaker verification, speech recognition and speech synthesis applications. Current speaker verification and recognition systems are limited by their lack of robustness against intra-speaker variability. This variability can cause unacceptably high error rates, which limits the commercial viability of such systems. To this point, intra-speaker variability has been treated largely as a random process, to be overcome by the use of better statistical techniques. A large body of research shows however, that many within-speaker variations in the voice can be traced to changes in the speaker's psychological and physiological state. An improved knowledge of the mechanisms of such speaker-state changes, in conjunction with a suitable database of speaker-state influences on speech, should lead to an improvement in the performance of speaker verification and recognition systems, both by a better selection of the acoustical parameters used and more appropriate speech elicitation techniques [1,2]. This research should also lead to an improvement of speech synthesis systems by providing a better understanding of how attitudinal and emotional information can be expressed in synthesized speech [3,4].

In order to produce experimental manipulations of affective speaker state in a controlled manner, a computer-aided speech recording tool was developed, in which the speaker is confronted with "computer tasks" meant to replicate natural working situations or situations from everyday life. Spontaneous speech as well as read speech are recorded while the speaker completes the different tasks. After each task, speakers are prompted to self-rate their emotional state, permitting an evaluation of the relative efficacy of different tasks in inducing the targeted emotional reactions. One of the aims of the current work is to examine whether ASV algorithms that are trained on emotional speech are more accurate when applied to speech samples obtained under conditions of stress or emotional activation than those trained on neutral speech only. In addition, our task is to determine in greater detail than has been the case until now, how stress and emotion affect voice production, and, in consequence, acoustic voice and speech parameters. This is particularly true for the study of spontaneous speech which has been rarely studied in this area, most work having been conducted with the help of actor portrayals. The assumption is that, in the long run, a better understanding of which acoustic parameters show great intraspeaker variance combined with high interspeaker stability across different types of fluctuating emotional states in real-life conditions can help to improve the accuracy and robustness of ASV algorithms.

2. METHOD

2.1 Computer-aided tool for emotion induction

The program is composed of six tasks. The first four tasks are designed to affect the speaker's emotional state and thereby to induce involuntary variations in their vocal expression. These four tasks were presented in random order to each speaker. The last two tasks required the speakers to deliberately control their speaking style and to voluntarily produce emotional expressions.

Tracking Task. This task was designed to induce stress through heightened perceptual-motor demands. It consists of a game where the user has to follow or to avoid a moving target using the mouse to control a figure on the screen. In a "conducive" condition the player/speaker wins points when he comes close to a "friendly" moving figure on the screen. In an "obstructive" condition the player/speaker loses points when an "unfriendly" figure gets close to him. Two levels of difficulty were superimposed on those two conditions, resulting in four different experimental conditions.

Number Sequence. This task is designed to induce irritation/anger and satisfaction. The user of the program has to complete an easy logic test where the time he needs to give the...
right answers is indicative of his performance. The user is slowed down by apparent “computer problems” and gets unjustified bad feedback on his performance. The assumption is that this manipulation will produce irritation in the speaker. For another subset of number sequences - when no “computer problems” arise - the users receive a positive feedback stating that their performance was better than the performance of former participants.

**Logic Deductions plus Auditory Stimuli.** This task is designed to produce cognitive stress in the user. It consists of a split attention task simulating a context where a person has to work while being disturbed by other stimulation. Users are asked to perform a logical reasoning test and an auditory monitoring task at the same time. Concretely, they have to make logical deductions on the basis of given premises displayed on screen, and respond to a particular sound when it occurs while ignoring another sound.

**Public Speech.** In this classic anxiety-producing task the participant is asked to present a short speech on a given topic. A judge/observer is present, makes notes and evaluates the presentation of the participant. The participants are also asked to give a short presentation on another topic but without their speech being evaluated. This second task was expected to induce less anxiety while representing a task of equal difficulty (i.e. with same level of stress induced by the difficulty of the task).

**Velten.** This task is designed to induce positive and negative feelings in the users. The procedure used in this task is a validated emotion induction procedure. Participants are asked to read short statements expressing positive/happy ideas or feelings or negative/sad ideas or feelings with the instruction to put themselves into the corresponding state of mind as much as possible before reading the statements out loud.

**Acting.** The final task is designed to record a greater range of speech variations from the participants by asking them to speak as if they were expressing different states/emotions.

Descriptions of 12 situations are given to the speakers who are asked to imagine the situations as vividly as possible and then read four standard phrases as if they were experiencing and expressing the corresponding states/feelings.

### 2.2 Speech recording and verbal feeling report

Pop-up windows at the beginning and end of each task (sometimes also during the task) prompt the user to read out standard phrases and series of numbers while performing the task. In the “Velten” task speakers read statements from the screen. The public speaking task involves free speech. In the acting task, standard phrases are presented and have to be produced with different emotional tones. All speech recording is entirely controlled by the computer program which starts and stops sampling via the computer’s audio card. A high-quality condenser microphone built into a headset is used (keeping the distance from the mouth relatively constant).

After each task, participants are asked to self-rate their emotional state by choosing an emotion word (among a list) and an intensity level describing their present state. This self-assessment of the emotional reaction is used as a control to evaluate the relative efficacy of different tasks in inducing the targeted emotional reactions.

### 2.3 Physiological measures

Measures of the participants’ physiological reactions have been taken for a subset of French speakers. Skin conductance level, number of skin conductance responses, rate of change of finger skin temperature, average respiration rate, heart rate and heart rate variability were used as independent indicators of cognitive and emotional reactions to the induction procedures. Software for the semi-automatic analysis of large amounts of physiological data that has been developed in our laboratory will be applied to this section of data analysis.

### 2.4 Speakers

Using the computer induction tool, 110 male speakers were recorded (30 were native German, 20 native English, and 60 native French speakers). Each speaker produced about 100 sentences of read speech and several passages of spontaneous speech.

### 2.5 Experimental Speaker Verification System

In collaboration with Ensigma Ltd. we developed a speaker verification procedure, which uses “structured training” to construct 2 speaker models with the help of speaker-independent AVS software systems. Speech data from 9 speakers (3 English, 3 French and 3 German speakers) were used to pretest the ASV system modified to allow “structured training”. The verification study was carried out with 20 English, 29 French and 29 German speakers from the remaining pool of speakers.

Models were built in the following fashion: To build a “neutral” speaker model 34 standard sentences, recorded before and after each task, as well as 1 minute of fluent speech, in which the speaker could state his opinion about a given (non-emotional) topic, were used. The “emotional” speaker model is built from 12 standard sentences recorded during the logic deduction task, 10 standard sentences recorded during the tracking task and 1 minute of speech recorded in the presence of an observer. This selection procedure ensured that the training data for both models, the emotional and the neutral speaker model, have similar duration and verbal content.
For verification, each model (neutral and emotional) for each speaker was scored against all those mixed-emotion test files that had not been used for model building. Closed set normalization was used. Both models are tested with the same set of utterances, including 36 sentences form the acting task, 4 explanations given in the number sequence task, and 12 sentences from the “Velten” task.

2.6 Acoustic analysis

An extensive set of automatic acoustic analysis routines was developed covering both established acoustic parameters that have been implicated in the expression of emotion in the voice and a number of promising new parameters.

3. RESULTS TO DATE

This contribution being a progress report, only some illustrative results can be presented. The extensive acoustic analyses and the fine-tuning of the verification procedures are ongoing. In this report we focus on the 30 German speakers for whom some of the analyses have been completed.

Figure 1: DET plot for the German mixed-emotion test samples for 29 speakers compared to models based on training for neutral (upper line) and mixed-emotion (lower line) speech samples.

Figure 1 shows the DET plot for a first verification study with the German speakers conducted in collaboration with Ensigma. The graph shows that, while the size of the effect seems relatively small, the mixed-emotion model is consistently more accurate with a lower error rate across the whole range. Clearly, these first results need to be examined in more detail and compared to similar analyses for the French and English speakers. One way to examine the statistical significance of the difference between neutral and mixed-emotion models is to compare the recognition scores for speakers across the different test samples. This is shown in Figure 2, demonstrating a significant improvement in the size of the recognition score in the case of the mixed-emotion model.

Figure 2: Difference in raw recognition scores for neutral and mixed-emotion models (t-test, p = .0036, N = 20)

However, it is felt that what is more important than the overall results of the verification, as expressed in DET plots or Equal Error rates, is a more fundamental understanding why and in which cases the use of a model trained on mixed-emotion speech samples is likely to provide superior results. Theoretically this should be the case for speakers who show a large variability of their vocal features depending on their respective attitudinal or emotional state. Whereas speakers with a rather limited amount of intra-speaker variability over speech situations should be recognized reasonably well by the classic procedure, based on neutral test phrases or lists of numbers, highly expressive speakers, showing large excursions of the values for different acoustic parameters, should be more difficult to recognize by neutral models.

Research on the acoustical correlates of emotional speech has rarely dealt with such individual differences in voice variability since most of the studies have used actors to portray emotional stimuli or have only studied a very small number of real-life speech samples by individual speakers (like the famous Hindenburg zeppelin crash reporter, [5]). Differences between actors are often considered to be due to different acting schools or styles.

However, there are indications in the literature that individuals vary greatly in their vocal reaction to affective stimuli including stress. Thus, Helfrich, Standke, & Scherer [6] showed significant differences between high and low depressive speakers in the acoustic effects of different kinds of antidepressive drugs. Tolkmit & Scherer [7] found that F0 floor (lowest 5% of F0 values) was higher under both cognitive and emotional stress for high anxiety and anxiety denying subjects. Low anxiety subjects showed an opposite change in F0 floor. Most importantly, the authors showed that de-peripheralization of vowels (as measured by the distance between observed and normative formant frequencies) occurred only for female (rather than male) subjects, showing strong interaction effects with type of stressor and personality trait. Female subjects characterized by a strong tendency to deny anxiety showed a strong tendency to de-peripheralize vowels under emotional stress but to peripheralize them under cognitive stress. More recently, Karlson and her associates in the Verivox project [8] measured the influence of computer game experience on stress and consequent voice changes. 41 of the 50 recorded speakers were divided into two groups: habitual computer game players (19) and non-players (22). One typical utterance (digit sequence) was acoustically
analyzed. The players indicated a lower mean stress level than the non-players, possibly because of greater experience with stressful interaction with the computer. As expected, this difference affected the size of the observed vocal changes. While non-players lengthened the utterance during a cognitive stress task, the players shortened it. While F0 increased by about 3% for the non-players under cognitive stress, no difference was found for the players (see [8,9] for further details).

While these differences are probably partly due to psychological differences (cognitive style, personality, learning, etc.) in the reaction to stress- and emotion-inducing experiences, it cannot be excluded that physiological factors, directly linked to the voice production process, are also involved. As mentioned above, such differences should be investigated more systematically and taken into account in speech technology development.

In the EMOVOX project we intend to investigate these individual differences in vocal response to stress and emotion by looking at the physiological measurements that were obtained for part of the speakers, and, in particular, at the acoustical parameters extracted from the different speech samples. For example, Figure 3 shows the differences between speakers and the intraspeaker variability over the different emotions in the "Acting" task for the proportion of spectral energy below 500Hz. The median for this parameter correlates with r = -.47 (p < .05, N = 19) with the mixed-emotion model recognition scores generated by the verification algorithm. Thus, speakers are better significantly recognized if they have a higher proportion of energy in the upper frequencies of the spectrum. It should be noted, however, that the correlation for the neutral model points in the same direction, without reaching significance. This may indicate that using structured training with emotional speech samples will particularly boost recognition for speakers with a high proportion of spectral energy above 500Hz.

4. CONCLUSION

While the evidence is so far inconclusive, data are accumulating that show that it is worth to consider emotional and attitudinal speech (see [10]) as a serious issue for speech technology research. It is theoretically reasonable to assume that training of ASV algorithms on a wider range of speech samples, including stress and different types of emotional arousal, rather than the standard neutral samples, is a promising way of improving recognition accuracy. In order to benefit more fully from taking such factors into account, it may be useful to investigate the role of individual differences in vocal reactivity and recognizability.

5. REFERENCES