A STATISTICAL APPROACH TO ASSESSING SPEECH AND VOICE VARIABILITY IN SPEAKER VERIFICATION

K. R. Scherer, D. Grandjean, T. Johnstone, G. Klasmeyer, and T. Bänziger

Department of Psychology
University of Geneva, Switzerland
Klaus.Scherer@pse.unige.ch

ABSTRACT

Voice and speech parameters for a single speaker vary widely over different contexts, in particular in situations in which speakers are affected by stress or emotion or in which speech styles are used strategically. This high degree of intra-speaker variability presents a major challenge for speaker verification systems. Based on a large-scale study in which different kinds of affective states were induced in over 100 speakers from three language groups, we use a statistical approach to identify speech and voice parameters that are likely to strongly vary as a function of the respective situation and affective state as well as those that tend to remain relatively stable. In addition, we evaluate the latter with respect to their potential to differentiate individual speakers.

1. INTRODUCTION

The starting point for this research project, which is a follow-up study of a collaboration between different European speech technology and phonetics labs [1,2,3], financed by the Swiss National Research Fund, was the observation that the performance, and hence the acceptability, of Automatic Speaker Verification (ASV) systems suffers from a sizeable drop in recognition accuracy when the speaker to be verified is affected by stress and emotion. Two major approaches were suggested to address this issue, 1) structural or mixed training of ASV systems with neutral and emotional speech and 2) identifying acoustic parameters that show both minimal intra-speaker variation over different psychological states and maximal inter-speaker variation, allowing recognition of individual speakers.

1.1. Training commercial ASV systems on both neutral and affectively charged speech samples

The assumption underlying this approach was that the models for the target speakers might be more stable if they were trained on both habitual speech patterns of the speakers and samples taken from situations in which speakers were under stress or experiencing a particular emotion. As described in detail elsewhere [4], we selected a sample of both neutral speech samples and emotionally charged speech samples for speakers in 3 languages. These samples served as material in a collaborative study with Ensigma Ltd., a British speech technology laboratory marketing ASV software. Under contract, Ensigma used a subset of the neutral and emotional samples to train either neutral or mixed (neutral plus emotional) models. The remaining neutral and emotional samples were then used to test differences in recognition accuracy for both types of training.

Using this approach, small but reliable improvement in overall performance was achieved. As expected, much of the improvement is obtained through a reduction in the number of false rejections. While one could probably augment the effect of structural or mixed training by more careful selection of the number and type of speech samples, it is not certain that maximal recognition accuracy can be achieved in this manner. Thus, it seems of interest to explore the second approach – trying to find acoustic parameters that are less subject to the influence of stress or emotional state on the voice while varying widely over different speakers.

1.2. Determining the relative variability of voice parameters over speakers and situations

We argue that performance of ASV systems can be improved by selecting a set of acoustic parameters which show both minimal intra-speaker variation and maximal inter-speaker variation and by building speaker models that are based on such optimal parameter sets (or to differentially weight parameters in a standard set). Such a procedure should reduce both false rejections and false acceptances. We have adopted an approach to this issue that is informed by the statistical notion of variance explained as generally used in experimental psychology. When psychologists manipulate variables as factors in an experimental design, they seek to determine how much of the total variance that is empirically observed in the data can be attributed to these factors and their higher-order interactions. The variance explained by the manipulated factors is compared to the variance between or within subjects, considered as error, and the significance of the contribution of a factor or an interaction term is evaluated by test statistics derived from normal distribution of error assumptions.

In the case of ASV, the situation is different in that the variation of acoustic parameters over speakers is not the error but rather the very basis of the automatic recognition of individual speakers. The more speakers' voices vary on acoustic parameters, the better they can be discriminated by an ASV system. Until now, this issue has been somewhat neglected by the developers of ASV systems since the large majority of ASV solutions are based on cepstral measures that attempt to capture speaker-specific energy distribution in the spectrum. Thus, the issue of the differential utility of widely different acoustic parameters does not pose itself.

However, it might be worth investigating a more general approach to the ASV problem by measuring a larger number of acoustic parameters. It is been observed in the literature that there are major differences in the variability of acoustic parameters over speakers. For example, whereas there is a fairly narrow range of F0 for men and women, voice quality or timbre varies more widely and is generally considered to constitute the basis for distinguishing individual speakers. To our knowledge, there have been few attempts to systematically determine the differential variance of different types of acoustic parameters over speakers (but see [5]) and thereby to determine their utility for ASR. This has been the aim of the analyses to be reported below.

Concretely, one would want to use acoustic markers for ASV that are 1) maximally different across sets of speakers (and languages or dialects), thus constituting a uniqueness dimension, and 2) robust with respect to the effect of transitory changes in speakers' affect states, attitudes and illocutionary stances as induced by internal factors (e.g., fatigue) or environmental influences (e.g., stress or emotion induction through specific situations or events), constituting a stability dimension. With respect to 2), prior research has shown that acoustic parameters vary with respect to the extent that they are affected by such changes in speaker state [4,6,7]. In consequence, one type of analysis of the large data set obtained in this project was to determine the relative promise of different parameters in this context.

We decided to give priority to the stability dimension (point 2. above). This is justified by the fact that ASV applications need to be based on a limited number of templates for a specific speaker and that these are usually obtained in relatively neutral situations. Thus, it is important to select acoustic markers of individual identity that show little effect to state changes in speakers. Once one has controlled for changeability across situations and states, one can evaluate parameters for their distinctiveness.

2. METHOD

2.1 Computer-aided tool for the induction of affective states in different situational contexts

A computer program was developed to record speech under various conditions designed to induce emotional variations in speech. The aim was to produce the same kind of variations that are likely to occur in everyday situations and to affect speaker verification systems in real settings. The speaker is confronted with “computer tasks” meant to replicate natural working situations or situations from everyday life. The program elicits various kinds of speech material. Spontaneous speech as well as read speech is recorded while the speaker completes the different emotion inducing tasks. The first four tasks are designed to affect the speaker's emotional state and thereby to induce involuntary variations in their vocal expression. The last two tasks require the speakers to deliberately control the way they speak and to produce voluntary emotional expressions. The first four tasks, designed to affect the speaker's emotional state and thereby to induce involuntary variations in their vocal expression, were presented in random order to each speaker. The last two tasks, requiring the speakers to deliberately control their speaking style and to voluntarily produce emotional expressions always appeared in the same order. In what follows, the six tasks are briefly described.

Tracking Task. Designed to induce stress through heightened perceptual-motor demands, the task is presented as a game where the user has to follow or to avoid a moving target using the mouse to control a figure on the screen. Two levels of difficulty were superimposed on a success/failure level, resulting in four different experimental conditions.

Number Sequence Task. Designed to induce irritation/anger and satisfaction, the task requires the user to complete an easy arithmetic test where the time he needs to give the right answers is indicative of his performance. The user is asked to complete number sequences by choosing the right number from a list. For a subset of number sequences the user is slowed down by apparent "computer problems" (the cursor disappears, on-screen buttons are inactive, numbers the user moves jump back to their original position...) and gets an - obviously unjustified - bad feedback on his performance.

Logical Deduction/Auditory Monitoring Task. Designed to produce cognitive load and psychological stress in the user, it consists of a split attention task simulating a context where a person has to work while being disturbed by another 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 Task. Designed to induce anxious stress, the speaker is asked to present a short speech on a given topic. Conforming to a social anxiety induction procedure used in emotion psychology research, a judge/observer is present, makes notes and evaluates the presentation of the participant.

Velten Emotion Induction. Designed to induce positive and negative feelings in the users, this established and frequently validated induction procedure requires speakers 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 Task. Designed to record a great range of speech variations, speakers are required to vocally express or portray a range of 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.

Neutral state. In addition to these six computer-administered tasks, speech samples were obtained during a break in the procedure, designed to allow speakers to rest and collect voice samples for a neutral, relaxed psychological state.

2.2. Speech recording

Pop-up windows at the beginning and end of each task as well as at different times during the task prompt the user to utter standard phrases and series of numbers while performing the

task (for example: "This is task number 345629"; with the number changing for each subtask). Thus, while part of the utterance is standardized, another part varies across subtasks. 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).

2.3. Speakers

Using the computer-induction tool, 103 male speakers were recorded (27 native German, 17 native English, and 59 native French speakers). Each speaker produced about 100 sentences of read speech and several passages of spontaneous speech.

2.4. Acoustic analysis

An extensive set of automatic acoustic analysis routines was developed by one of the authors (GK), covering both established acoustic parameters that have been implicated in the expression of stress and emotion in the voice, as well as a number of more rarely used parameters.

2.5. Statistical analysis

For the analyses reported in this paper, we retained only the standardized speech samples as described above. Using the complete data matrix obtained in the current study (three groups of speakers with different languages, 7 different tasks with different sets of conditions, and several repetitions of utterances within task conditions), we ran univariate ANOVAs for all acoustic parameters measured, using a design of speakers nested within language as a random factor and a variable representing task by condition (the string of different speech situations constituted by the conditions across tasks) as a fixed factor. Language and the interaction of language with the task/condition factor were also included .

For this analysis we chose a hierarchical partitioning of the effect (as measured by eta^2; proportion of variance explained), using Type 1 sums of squares. This type of analysis requires that one specifies the order in which the factors are introduced into the model. The analysis then determines for each factor the proportion of the variance it explains out of the residue of variance that is not explained by the factor entered previously. In line with the priority defined above, we entered the task/condition factor first, followed by the language factor, followed by task/condition*language interaction, and then the subject/speaker factor. This allows to first determine the proportion of the variance explained by the speaker state changes induced by our experimental manipulation (which reflects the relative stability or robustness of these parameters) and then determine how much of the remaining variance can be accounted for by individual differences. In addition, by adjusting the eta^2 values to the amount of variance remaining after the variance explained by the preceding factor is accounted for, we estimate the relative contribution of the situational and the individual difference factors to the total variance.

3. RESULTS

Figure 1 shows a graphical representation of the proportion of the variance in individual speaker differences explained by the acoustic parameters measured in this study, sorted by size. This allows to identify the acoustic parameters for which a high proportion of variance is accounted (Explained Variance, EV) for by speaker differences after the variance explained by situation has been removed.

Figure 1 shows an almost perfect linearity with respect to the power of different acoustic parameters to explain individual voice variability. Contrary to what one might have thought, there is no clear cut-off point that would allow establishing a straightforward classification of useful and useless acoustic parameters for the task of automatic speaker recognition. Obviously, then, one needs to employ more sophisticated statistical tools in trying to identify the variables that are likely to distinguish speakers without being too much subject to emotional and situational variation.

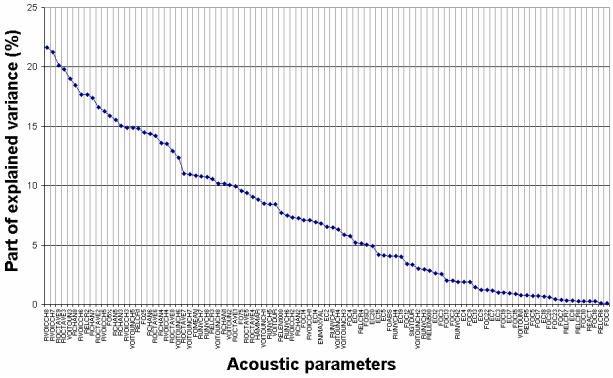

Figure 1: Proportion of variance explained by the speaker factor for 99 acoustic parameters

We first ran a series of Discriminant Analyses to test how well sets of acoustic parameters differing in the number of predictors used allow automatically classifying speakers, i.e., determine the identity of individual speakers. The percentage of cases correctly classified on the basis of these parameter sets are shown in Table 1. The High EV (Low EV) grouping corresponds to the acoustic parameters for which the speaker factor explained most (least) variance by subject factor (top and bottom of the distribution shown in Fig. 1). The discriminant analysis functions were crossvalidated by calculating functions on a randomly chosen half of the data (High and Low EV) and then testing their stability on the other half (CrossHigh and CrossLow EV).

Table 1: Percentage of speakers correctly classified by acoustic parameters.

Nb of acoustic parameters		5	10	15	20	25	30	35	40	45
Percentage of speakers correctly classified (N=103)	High EV	13.3	23.5	35.5	45.2	50.8	54.1	56.1	60.2	62.8
	Low EV	3.2	5.2	7.9	10	12.6	21	33.3	39.3	44.3
	CrossHigh EV	13.1	22.9	34.4	44.1	49.5	52.7	54.6	58.3	60.9
	CrossLow EV	2.7	4.4	6.5	8.3	10.2	17.5	28.9	34.5	39.5

We decided to use a cut-off level of 25 acoustic parameters which correctly discriminate approx. 50% of the speakers for further analysis.
We compared the results of the series of discriminant analyses that attempt to classify speakers on the basis of acoustic

parameters explaining a high percentage of individual variance once situational variance is factored out, by attempting to discriminate the experimentally manipulated situational and emotional states, the task/condition factor, with the same acoustic parameters as in the earlier analyses. The results are shown in Fig. 2.

The data illustrated in Fig. 2 demonstrate the incremental utility of acoustic parameters that explain a high percentage of individual variance once situational variance is factored out. These parameters are not redundant, adding additional ones will cumulatively increase the classification accuracy. Conversely, there is a floor effect for the prediction of situational and emotional states, an increase in the number of predictors selected to identify individuals does not improve classification success.

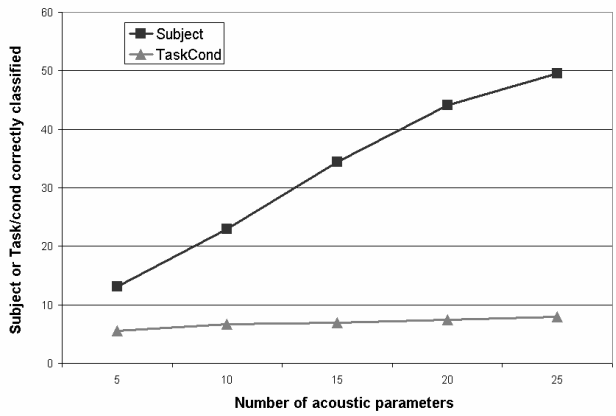

Figure 2: Comparison of the capacity to discriminate speakers (subject factor, N=103) and experimental task conditions (N=25) with the same sets of acoustic parameters.

Examining the intercorrelations between the different acoustic variables it became quickly apparent, as one expects from the literature, that there was a high level of collinearity in the parameter set. We therefore conducted a series of factor analyses of the large number of spectral parameters extracted and constructed on this basis, as well as by examining the correlation matrices, a set of aggregated acoustic variables reflecting the energy distribution in the spectrum that capture the unique contribution of specific acoustic dimensions to the explanation of the variance. These aggregated variables were subjected to ANOVAs, assessing in particular the relative amount of variance explained by the speaker factor compared to the situation factor as measured by effect size (eta^2). In order to summarize the relative importance of the respective parameter to classify individuals, we computed a *stability quotient* by dividing the eta^2 for situation by the percentage of variance explained by the speaker factor once the situation variance had been controlled for. In trying to decide which aggregate variable was a good predictor for individual speaker differences, we used a cut-off point of .10 for this quotient.

Based on the statistical evidence produced by this procedure, we concluded that F0 floor (the lower 5% of the values in the F0 distribution) and the relative energy (for voiced speech segments) between .125-.5 seems to predict relatively more individual speaker variation than emotion-induced variation. In contrast, the relative energy (for voiced speech segments) between .5-1.6 kHz seems to be relatively more determined by emotional and situational variation than by inter-speaker variation. These results hold across all three languages studied.

4. CONCLUSION

In this contribution we suggest to rethink the acoustic fundamentals of standard ASV procedures. We propose that a detailed acoustic analysis of speech samples for a large number of speakers collected under a wide variety of situations (likely to produce varied affective states), can provide important information as to the capacity of different parameters to mark speaker individuality and to resist the effect of changing psychological and physiological states. While the findings obtained in the large-scale study reported here need further replication, the statistical stability of the results suggest that the general approach is promising. As would be expected on the basis of past work, voice variability that can serve as a signature for speaker identity is mostly to be found in voice quality, more specifically, in energy distribution across the spectrum. Importantly, the physiological changes associated with stress and emotion also primarily affect voice quality as measured by energy distribution in the spectrum. Fortunately there seems to be a reliable difference with respect to the region of the spectrum (for voiced speech segments) that is differentially affected by situational affective change and individual speaker characteristics. While the present results require replication, we believe that the underlying research strategy has interesting implications for speech technology applications.

5. REFERENCES

[1] I. Karlsson, T. Bänziger, J. Dankovicova, T. Johnstone, J. Lindberg, H. Melin, F. Nolan, and K.R. Scherer, "Speaker verification with elicited speaking styles in the VeriVox project," Speech Communication, vol. 31, no. 2-3, June, pp. 121-129, 2000.

[2] K.R. Scherer, T. Johnstone, and T. Bänziger, "Automatic verification of emotionally stressed speakers: The problem of individual differences," Proc. of SPECOM98, St. Petersburg, 1998.

[3] K.R. Scherer, T. Johnstone, G. Klasmeyer, and T. Bänziger, "Can automatic speaker verification be improved by training the algorithms on emotional speech?," Proc. ICSLP2000, Beijing, 2000.

[4] G. Klasmeyer, T. Johnstone, T. Bänziger, C. Sappok, and K.R. Scherer, "Emotional voice variability in speaker verification," Proc. ISCA Workshop Speech Emotion, Belfast, 2000.

[5] K. Johnson and J. Mullenix, Eds., Talker variability in speech processing, San Diego : Academic Press, 1997.

[6] R. Banse, and K.R. Scherer. "Acoustic profiles in vocal emotion expression," J. Pers. Soc. Psychol., vol. 70, no. 3, Mar., pp. 614-636, 1996.

[7] T. Johnstone, and K.R. Scherer, "Vocal communication of emotion" in Handbook of Emotions, Second Edition, M. Lewis and J. Haviland-Jones, Eds. New York: Guilford Press, 2000. pp. 220-235.