ADDING THE AFFECTIVE DIMENSION: A NEW LOOK IN SPEECH ANALYSIS AND SYNTHESIS

Klaus R. Scherer
scherer@uni2a.unige.ch

University of Geneva

ABSTRACT

This introduction to a special session on "Emotion in recognition and synthesis" highlights the need to understand the effects of affective speaker states on voice and speech on a psychophysiological level. It is argued that major advances in speaker verification, speech recognition, and natural-sounding speech synthesis depend on increases in our knowledge of the mechanisms underlying voice and speech production under emotional arousal or other attitudinal states, as well as on a more adequate understanding of listener decoding of affect from vocal quality. A brief review of the current state of the art is provided.

1. INTRODUCTION

While recent years have seen major advances with respect to speaker/speech recognition and synthesis, a number of issues remain unresolved. One of these is the variability in the speech signal which is due to emotional and attitudinal speaker states. Since affective arousal has a powerful effect on the voice, there are major changes in the acoustic parameters of speech accompanying different affect states. This is the case even for emotional states of relatively low intensity as they occur in everyday life. This affect-driven variability in the human voice is a major problem for accurate speaker verification under a variety of everyday situations and moods and for adequate speech recognition in non-neutral interactions. On the other hand, much of speech synthesis is flawed by the lack of appropriate affective variation in prosody and voice quality which seems to be required for both intelligibility and acceptability.

Many speech technology researchers have tended to downplay this problem and to assume that consistent improvement in mathematical algorithms would solve the problem. So far, this optimistic attitude has not paid off and it does not seem to harsh an analysis to claim that after impressive initial achievements, speech technologies have shown improvements in a rather asymptotic fashion, rendering many of them less than ideal for robust every-day application in real-life contexts other than simple messaging systems.

In setting the stage for this special session I will argue that tangible improvement in speech technology will require much greater attention to speaker affect and attitude effects than has been customary in this area. In the case of altered speaker states, the speech signal is not only affected by biophysiological push factors but also by prosody and a multitude of socio-normative, situationally bounded pull factors [1, 2]. In addition, past research has shown very significant individual differences in the vocal expression of emotions and attitudes. I will argue that, given the complexity of the determinant factors, progress in the ability of speech technologies to deal with multiple speaker states will be constrained by our ability to plot speaker-specific attractor spaces for vocal expression during emotionally charged episodes.

In the following, I will briefly summarize the literature, review some recent findings of a major study on vocal encoding and decoding of emotion, and address issues pertinent to recognition and synthesis.

2. REVIEW OF THE LITERATURE

Decoding: Typically, in studies of lay judges' ability to recognize emotions from purely vocal stimuli, an accuracy of about 60% is found [2]. This greatly exceeds what one would expect to obtain if the listener judgments were based exclusively on guessing, i.e. on chance (approximately 12%). The degree of recognition accuracy is impressive given that some of the studies included emotions such as love, pride, or jealousy which are not part of the set of basic or fundamental emotions (e.g. anger, joy, sadness, fear).

Sadness and anger are best recognized, followed by fear and joy. Disgust is the worst, with the accuracy barely above chance. The data show the need to analyze the recognizability of different emotions separately using confusion matrices, as errors are not randomly distributed and as the patterns of misidentification provide important information on the judgment process.

Encoding: Given the difficulty of inducing or observing naturally occurring vocal expressions of emotion, many researchers in this area have used actors as subjects, asking them to vocally portray different emotions, and have analyzed the acoustic features of the recorded portrayals. The state of the evidence up to 1995 has been summarized as follows [2]:

Anger: Anger is vocally expressed by an increase in mean F0 and mean intensity. Some studies, which may have been measuring "hot" rather than "cold" anger, also claim higher F0 variability and a wider range of F0. Further anger signs seem to be increases in high frequency energy and downward directed F0 contours. The rate of articulation usually increases in anger.

Fear: One expects a very high arousal level for fear. Consistent with this hypothesis, the data show increases in mean F0, in F0 range, and in high frequency energy. Rate of articulation increases.

In some studies, higher mean F0 is also reported for the weaker forms of fear (i.e. worry or anxiety).

Sadness: In sadness, mean F0, F0 range, and mean intensity all decrease, and F0 contours are generally downward directed. High frequency energy and rate of articulation decrease.

Joy: Studies consistently show increases in mean F0, F0 range, F0 variability, and mean intensity. There is some evidence for an increase in high frequency energy and in rate of articulation.

Whenever the findings in the literature converge, they seem to be due to the acoustic effects of sympathetic autonomic arousal. There has been relatively little evidence for the vocal differentiation of individual emotions on other dimensions such as valence. However, since judges are able to recognize the individual emotions on the basis of vocal cues alone, there must be acoustic characteristics that differentiate the various emotions in addition to indicating arousal. The difficulties of mapping the vocal-acoustic bases of emotional expression more clearly can be related to the fact that only a limited set of acoustic variables had been typically measured and that research procedures have rarely been guided by detailed theoretical hypotheses concerning the underlying mechanisms.

3. RECENT THEORY AND EVIDENCE

Based on a comprehensive theory of emotion, the present author has presented a set of precise predictions on the acoustic profiles of 14 major emotions [3]. Our group has tested these predictions in the context of a large-scale study on the expression of emotion in multiple communication modalities, [4]. 12 professional actors were asked to portray 14 emotions varying in intensity and valence or quality. The results on decoding replicate and extend earlier findings demonstrating the ability of judges to infer vocally expressed emotions for a large number of emotions (approx. 50% accuracy compared to 7% chance expectation). Consistently found differences in the recognizability of different emotions are also replicated.

A qualitative analysis of the confusion matrices yielded three dimensions of similarity: quality, intensity, and valence. The most obvious dimension of similarity is the quality of an emotion. Emotion pairs like Hot Anger and Cold Anger, Sadness and Despair, Anxiety and Panic Fear are similar in quality and differ mainly in intensity. A second dimension of similarity is intensity. For example, Elation was relatively often confused with Despair, Hot Anger, and Panic Fear, which differ strongly in quality, but are similar in intensity. The third dimension of similarity is the valence dimension. Positive emotions are more likely to be confused with other positive emotions than with negative emotions. If the three dimensions of similarity accounted for all errors, one would expect approximately symmetric confusions between emotions. However, this is not always the case. For example, there is substantial confusion of the Elation portrayals with Hot Anger, Panic Fear and Despair, but there are virtually no confusions of stimuli belonging to these three categories with Elation. One possible explanation for this finding may be an emotion specific "typicality" of acoustic features. That is, some emotions (e.g., Hot Anger) may be characterized by a very typical configuration of acoustic features, which are easy to identify. In this case, the underlying recognition mechanism is probably a prototype based top-down process. Other emotions like Elation, may lack "typicality". Decoders confronted with a display of elation may have to analyze the acoustic pattern in a "piecemeal" or bottom-up fashion, and may be easier mislead by prominent features like high intensity, which in the case of elation makes the stimulus similar to Hot Anger or Despair.

A total of 224 different portrayals, 16 per emotion category, were subjected to digital acoustic analysis to obtain profiles of vocal parameters for different emotions, using a larger set of acoustic variables than is normally employed in this research area. The data provide first indications that vocal parameters not only index the degree of intensity typical for different emotions but also differentiate valence or quality aspects. In particular, the data are used to test the theoretical predictions on vocal patterning based on the component process model of emotion [3]. While most hypotheses are supported, some need to be revised on the basis of the empirical evidence.

To test to what extent the acoustical parameters analyzed in this study allow correct emotion classification by machine, a jack-knifing procedure was performed. For each portrayal, the sum of the squared differences between the 29 individual acoustical parameter values and the mean profiles of the 14 emotions were calculated (for each comparison, mean profiles were always calculated without using the portrayal to classify). Each stimulus was then classified into the emotion category for which the sum of squared differences was minimal. A simple genetic algorithm was implemented to find a subset of parameters with optimal classification results. In this method, five parameters out of 29 are selected randomly, and their classification performance in the jack-knifing procedure is tested. In the next step, the selection is modified randomly by choosing or excluding five out of the 29 parameters, and tested again. If the performance is improved by the modification, the new combination is retained and becomes the basis for new random changes. The number of modifications is gradually reduced to allow for the identification of a local maximum. After 150 loops only three parameters are modified, after 300 loops two, after 400 one. After 500 loops the process is stopped. The results of the runs were rank-ordered by goodness-of-fit. After about 100 runs of this algorithm the performance tended to converge with respect to both hit rate and selected parameters. The best solution produced an overall hit rate of 40.4% (as compared to 7% expected by chance). The best performing subset of 16 of the total set of 29 acoustic parameters was the following: Fundamental frequency: Mean, standard deviation, 25th percentile, 75th percentile; Energy: Mean; Speech rate: duration of articulation periods, Bands in the voiced long term average spectrum: 125-200 Hz, 200-300 Hz, 500-600 Hz, 1000-1600 Hz, 5000-8000 Hz; Hammarberg index; slope of spectral energy above 1000 Hz; proportion of voiced energy up to 1000 Hz. Bands in the unvoiced long term average spectrum: 125-250 Hz, 5000-8000 Hz.

Because in this research both encoding and decoding were studied in parallel it was possible to regress the judges' emotion inferences on the various acoustic variables in order to derive first hypotheses on the use of these parameters in the judges' inference processes.

The highly significant results showed that a sizable proportion of the variance are explained by a set of about 9-10 variables, demonstrating that it is possible to determine the nature of the vocal cues that judges use in identifying emotional speaker state from vocal information. The comparison between the performance of human judges with statistical classification routines provides a promising approach to elucidate the inference mechanism at work by optimizing the selection and combination of acoustic parameters used by human judges'. In the present study this approach yielded a powerful result: Not only are the hit rates for correct recognition very similar, but, more importantly, there is a remarkable resemblance between the error patterns in the confusion matrices produced by humans and machine algorithms respectively. If these results can be replicated in future work, the importance of the 16 acoustic cues, found to be optimal in the jack-knifing procedure, would be underlined. While in the present case rather simplistic cue combination rules were used for the inference model one could imagine the development of much more sophisticated tools, e.g. as developed in artificial intelligence work, in this domain.

I suggest that the combination of theoretical predictions with the results from this massive data set provide a solid basis for further research on the effects of emotional speaker state with respect to speaker verification as well as speech recognition and synthesis.

4. SYNTHESIZING SPEAKER EMOTION AND ATTITUDE

Early studies showed the strong effects of synthetic amplitude variation, pitch level, contour and variation, tempo, envelope, harmonic richness, tonality, and rhythm on emotion attributions to sentence-like sound sequences and musical melodies [5,6]. Computer-based copy synthesis (or resynthesis) techniques have allowed researchers to systematically change different cues via digital manipulation of natural voices. In a series of studies conducted in our laboratory, the effects of F0 level, contour variability and range, intensity, duration, and accent structure of real utterances on emotion and attitude judgments have been systematically assessed [7,8]. The results showed strong direct effects for all of the variables manipulated on these ratings. Relatively few effects due to interactions between the manipulated variables were found. This implies that the synthesized variables independently influenced judges' ratings. Only very minor effects for speaker and utterance content were found, indicating that the results are likely to generalize over different speakers and utterances.

Of all variables studied, F0 range had the most powerful effect on judgments. Narrow F0 range was seen as a sign of sadness or of absence of specific speaker attitudes. Wide F0 range was consistently judged as expressing high arousal, producing attributions of strong negative emotions such as annoyance or anger, or for the presence of strongly developed speaker attitudes such as involvement, reproach, or emphatic stress. Furthermore, the data supported the hypothesis, derived from an earlier study by our group [9] that these effects should be continuous, i.e. yielding a near-linear relationship between the size of F0 range and the strength of emotion attribution. High intensity was interpreted in terms of negative affects or aggressive speaker attitudes. Short voiced segment duration (fast tempo) was correlated with inferences of joy, long duration (slow tempo) with inferences of sadness.

A recent review [10] shows the promise of synthesizing emotionally expressive speech and the number of pertinent studies seems to be on the increase (judging, for example, from the popularity of emotion-related contributions at last years ICPhS in Stockholm).

I firmly believe that further progress in synthesizing affective and prosodic speech qualities depends on the choice of the appropriate acoustic features. Such approaches need to take into account the intricate links between the function of the emotional state (including appraisal and action tendencies) and the corresponding physiological changes that directly affect the voice and speech mechanisms. Unfortunately, there has been little interchange between physiologically and acoustically oriented voice scientists and psychologists studying vocal emotion expression. Such links need to be established if we want to trace and model the mechanisms and processes whereby emotion-generated changes in the somatic and autonomic systems affect voice production (and thus ultimately the acoustic parameters we measure and synthesize in the speech signal). The selection and definition of the acoustic parameters that are directly pertinent to affective speaker state is still in its early stages. Many of the parameters used, particularly those related to the energy distribution in the spectrum are only very first approximations in trying to get at emotion specific acoustic changes. Since there is little established knowledge with respect to the effects of physiological arousal on voice production and the consequent changes in the acoustic speech signal, the measures used are largely based on speculation or empirical approximation. In addition to refining the voice parameters, more effort needs to be expended on developing reliable quantitative parameters for the measurement of suprasegmental features of speech such as rhythm, accentuation, and intonation. While such parameters have been used only rarely in this research area, the results that do exist suggest that prosodic parameters may play a major role in vocal emotion differentiation [7,8]. Advances in measuring the pertinent differences in emotion specific voice and speech signals are likely to strongly improve the ability of statistical models to accurately discriminate various emotional states.

It is to be hoped that future research activity in this area will avoid some of the shortcomings of earlier research on emotion encoding such as a completely atheoretical approach and the concentration a very small number of so-called basic emotions.

5. CONCLUSIONS

In fact, one of the lessons learned by emotion psychologists is that affective and attitudinal states are much more varied and complex than is often assumed on the basic of simplistic models of fundamental, discrete emotions. While it is extremely important to take into account the evolutionary continuity of vocal affect signaling, using ethological research, we also need to be very sensitive to language-specific effects of prosody and the cultural embedding of vocal communication in humans. Given the intricate

interaction between biological push factors and socio-cultural and linguistic pull factors, accompanied by very powerful individual differences in emotion-antecedent appraisal and response patterning, a paradigm shift with respect to research in this area will be required in order to allow our knowledge to advance. I have suggested to view emotions as episodes of tightly synchronized functioning of the various organismic subsystems or, in the sense of dynamic systems theory, as transient attractors. A major task for future research will be the comprehensive analysis of these multi-system synchronisations, requiring complex multivariate time series measurements in different domains. Given the importance of individual differences, it will not be sufficient to identify a general map of emotion-specific attractors (and their vocal-acoustic concomitants) but we may well be required to map emotion-pertinent attractor spaces for individuals, taking into account specificities in modal response and transition patterns.

6. ACKNOWLEDGMENTS

Research reported in this paper has been supported by funding in the context of the ESPRIT-BRA VOX project and Swiss National Scientific Research Fund (project FNRS 1114-037504.93). The author would like to thank Tom Johnstone for valuable comments and suggestions.

7. REFERENCES

1. Kappas, A., Hess, U., and Scherer, K.R. “Voice and emotion”, In R.S. Feldman, and B. Rimé (Eds.), *Fundamentals of nonverbal behavior (pp. 200-238)*. Cambridge: Cambridge University Press, 1991

2. Scherer, K. R., “Expression of emotion in voice and music”, *J. Voice, 9(3), 1995, 235-248*.

3. Scherer, K. R. "Vocal Affect Expression: A review and a Model for Future Research," *Psych. Bull., 99, 1986, 143-165.*

4. Banse, R. and Scherer, K. R. "Acoustic profiles in vocal emotion expression," *J. Person. Social Psychol., 70, 1996, 614-636.*

5. Lieberman P., and Michaels S.B. “Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech,” *J. Acous. So. Ame., 34, 1962, 922-927.*

6. Scherer, K.R., and Oshinsky, J., “Cue utilization in emotion attribution from auditory stimuli”, *Motiv. Emot., 1, 1977, 331-346.*

7. Ladd, D., Silverman, K., Tolkmitt, F., Bergmann, G., and Scherer, K., “Evidence for the independent function of intonation contour type, voice quality, and F0 range in signalling speaker affect”, *J. Acoust. Soc. Amer., 78, 1985, 435-444*.

8. Tolkmitt, F., Bergmann, G., Goldbeck, Th., and Scherer, K.R., “Experimental studies on vocal communication”, In K.R. Scherer (Ed.), *Facets of emotion: Recent research. (pp. 119-138)*. Hillsdale, NJ: Lawrence Erlbaum, 1988.

9. Scherer, K.R., Ladd, D.R., and Silverman, K., “Vocal cues to speaker affect: Testing two models”, *J. Acoust. Society of America, 76, 1984, 1346-1356.*

10. Murray, I. R., and Arnott, J. L. "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoust. Soc. Amer., 93, 1993, 1097-1108.*