AUTOMATIC DETECTION OF STRESS IN SPEECH

H. J. Fell[1], J. MacAuslan[2]
[1]College of Computer and Information Science, Northeastern University, MA, USA
[2]Speech Technology and Applied Research, Lexington, MA USA

Abstract: **We have developed software based on the Stevens landmark theory to extract features in utterances in and adjacent to voiced regions. We then apply two statistical methods, closest-match (CM) and principal components analysis (PCA), to these features to classify utterances according to their emotional content. Using a subset of samples from the Actual Stress portion of the SUSAS database as a reference set, we automatically classify the emotional state of other samples with 75% accuracy, using CM either alone or with PCA and CM together. The accuracy apparently does not depend strongly on measurement errors or other small details of the present data, giving confidence that the results will be applicable to other data.**

Keywords : **automatic detection, emotion, speech, stress**

I. INTRODUCTION

If computers are to interact with humans in a natural way, they will need a speech interface that recognizes emotional as well as linguistic content of speech. Scherer *et al* [1998] argue that modeling of speaker states and emotions can improve the quality of automatic speech recognition, speech synthesis, and speaker verification and that such emotion effects are relatively robust to changes in the phonetic context. Imagine your computer responding with sympathy when you are sad, explaining things more simply when you are frustrated, or speaking calmly to you when you are stressed.

Speech scientists have been able to identify a number of acoustic speech parameters that correlate with the speaker's emotional state. Johnstone & Scherer [6] report that analysis of glottal opening and closing characteristics proved useful in interpreting the emotion-dependent characteristics of the acoustic waveform. Quast [10] identifies a number of parameters that appear to carry crucial information, e.g. location of the sentence foci, intensity values, relation of the fundamental frequencies (F_0) at the focus and ends of the sentence, speech rate, and spectral histogram.

There have been few attempts and limited success at actually recognizing and classifying affect in speech. Roy and Pentland [11] used six acoustic measurements (F_0 mean and variance, Energy variance and derivative, open quotient, and spectral tilt) to classify spoken sentences as approving or disapproving. They achieved 65% to 85% classification accuracy for speaker dependent, text independent data. Their results suggest that energy and F_0 statistics may be effectively used for automatic affect classification. Stolcke et al. [14] used prosodic cues as part of a statistical approach to model dialogue acts in conversational speech. They achieved a 71% accuracy in labeling act-like units such as statement, question, agreement, disagreement, and apology. Dellaert et al. [1] applied several statistical pattern recognition techniques to classify utterances according to their emotional content. For the purposes of classification they used only pitch information extracted from the utterances. They also introduced a spline approximation of the pitch contour to extract features. Their best method resulted in a 20.5% error rate in classifying four emotions: happiness, sadness, anger, fear. Human performance at the same task resulted in an 18% error rate.

We have had success in applying landmark detection coupled with Principal Component Analysis in detecting significant differences in the vocalizations of typically-developing and at-risk infants [2, 3, 4] and in detecting fatigue in adult speech [8]. Here, they apply similar techniques to classifying stress in speech.

II. THE DATA

We are using the Actual Speech Under Stress portion of the SUSAS (Speech Under Simulated and Actual Stress) database [5]. A common highly confusable vocabulary set of 35 aircraft communication words make up the database. All speech tokens were sampled using a 16-bit A/D converter at a sample rate of 8kHz. We are using samples recorded under four conditions: neutral - Neutral Speech, medst - low Dual-Tracking task stress, hist - high Dual-Tracking task stress, and scream - Scream Machine Roller Coaster stress. We have restricted this study to the four male speakers: m1, m2, m4, who have a General USA Accent; and m3: who has a Southern USA Accent.

We formed a base of features for classification using only the first sample of each of the 35 words for each speaker in each emotional state whenever such samples were present. Table 1 shows the number of words for each speaker/emotional-state used to create the base.

Table 1: Number of words used to create the base for classification

	neutral	medst	hist	scream
m1	35	35	35	29
m2	34	35	35	29
m3	34	35	35	23
m4	35	35	35	23

We then created test cases for classification using the second sample of each of the 35 words for each speaker in each emotional state whenever such samples were present. Table 2 shows the number of words for each speaker/emotional-state test case.

Table 2: Number of words per sample for the 16 test cases

	neutral	medst	hist	scream
m1	35	35	35	15
m2	8	35	35	24
m3	34	35	34	15
m4	35	35	35	2

III. Methodology

We listened to many words in the SUSAS Actual Stress database before attempting to perform automatic classification. One subjective impression was that the vowels were longer, relative to word duration, in the medst and hist words than in the corresponding neutral words. Another impression was that the consonants were clipped, shorter and less structured than their neutral correspondents. To model these impressions, we needed to extract more than pitch information.

Using software that we have developed [2, 3, 4] based on the Stevens landmark detection theory [7, 13] for the recognition of phonetic features in speech, we extracted measurements on twenty-five features from the ~35-word sets of speech samples. These served to summarize the speaker, state, and sample.

From Syllables:
Timing:
mean duration, mean duration of voicing, mean voiced fraction (i.e. mean of voiced duration/total duration), maximum and mean voice onset time (VOT), maximum and mean offset time, mean rate (i.e. mean of 1/duration), mean voiced rate (i.e. mean of 1/voiced duration).
Pitch (F_0):
median and mean F_0, fraction of syllables in which the pitch rises (falls) during the first half (second half) of the syllable.
Structure:
mean, median, and maximum number of landmarks per syllable.

From Words:
Pitch:
root mean square standard deviation of F_0, relative range of pitch (see below), 10^{th}, 50^{th}, and 90^{th} percentile value of the relative range, 10^{th}, 50^{th}, and 90^{th} percentile value (over all the words) of the "central" F_0 value, i.e., the value in the middle of the word.

The relative range of pitch is defined as the maximum (over each word) of the 90^{th} percentile values of the pitch, minus the minimum of the 10^{th} percentile values, divided by the median value (over the word). Thus, it is a non-negative number, and typically less than 1. We divide by the median F_0 so that the results are not strongly skewed for irrelevant reasons, such as a generally lower F_0 for men than women.

For each state, we normalized the four speakers' data by comparing their values for each of these features to the mean and standard deviation σ of all four in that state. Specifically, we subtracted the mean and then divided by a certain variability measure. This measure consists of σ and an *a priori* estimate of measurement error, combined in RSS (root sum-of-squares) fashion. Thus, for example, the squared measure for an F_0-related feature consists of the sum of the observed four-subject value of that feature's variance σ^2 plus $(5\text{ Hz})^2$, because 5 Hz represents an estimate of the irreducible measurement uncertainty for F_0. Such irreducible measurement uncertainties depend primarily on the recording environment or computational details (for F_0, at least).

Observe that this normalization process yields feature values of zero mean and approximately unit variance for the base cases. As 25-element vectors, then, the normalized base-case summaries have norm (Euclidean length) $\sim 25^{1/2}$.

When comparing one speaker/state/sample summary to another, we simply evaluate the RSS of the vector of differences in feature values. By construction, this also produces values $\sim 25^{1/2}$ to $50^{1/2}$ when comparing two base cases, and we might anticipate similar or even smaller results when comparing two samples from the same speaker and state. In fact, this was routinely observed.

To identify a state from a test set of speaker/state/sample, we hypothesize a state, normalize the corresponding summary using the mean and variability parameters for that state, and compare to each of the base cases of the state. Across all speakers and states defining the base, 16 summaries in all, the lowest RSS difference identifies the closest-matching, or CM, state (and, in principle, speaker).

An important refinement is available. Of the 16 subject/state normalized feature vectors that define the base, some linear combinations may be redundant. Eliminating these would improve the robustness of the results, because the redundant components would otherwise tend to model inappropriately small details of the data, i.e.,

"noise". Principal Components Analysis (PCA: equivalently, singular value decomposition, SVD) determines the extent to which this occurs among the set of vectors. In this case, the first three PC's accounted for 99% of the total variance, suggesting both a high degree of linear dependence and a high degree of linear predictability.

IV. Software and Algorithms

Our landmark detector is based on Stevens' acoustic model of speech production [13]. Central to this theory are *landmarks*, points of abrupt spectral change in an utterance around which listeners extract information about the underlying distinctive features. They mark perceptual foci and articulatory targets. Our program detects three types of landmarks:

glottis (+g, -g): marks the time when the vocal folds start and stop vibrating;
sonorant (+s, -s): marks sonorant consonantal closures and releases;
burst (+b, -b): aspiration/frication ends due to stop closures.

Our analysis is based on a low-resolution spectrogram. The SUSAS signals are sampled at 8 kHz and analyzed into a small number, nominally 32, of separate, frequency intervals of ~256 Hz each. An 8 kHz rate provides information only up to 4 kHz, but this is sufficiently high to include at least 3-4 formants for an adult and to show the distinction between voicing and other speech sounds: fricatives, stop releases, bursts, etc. (See Fig. 1.)

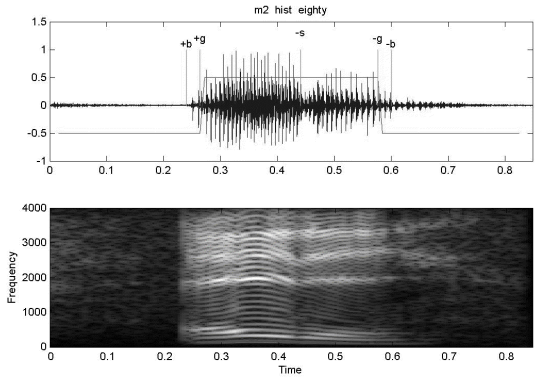

Figure 1: Waveform and Landmarks (top) and Spectrogram (bottom) of "eighty" as spoken by male 2 in high stress conditions.

To locate the landmarks, spectral intervals are grouped into six broad bands. An energy waveform is constructed in each of the six bands, the time derivative of the energy is computed, and peaks in the derivative are detected. These peaks thus represent times of abrupt spectral change in the six bands. Energy in bands 2 (1200 - 2500 Hz.) and 3 (1800 - 3500 Hz), e.g., provides evidence of voicing or, in some cases, of bursts. The distinction between these is readily made in the time domain (voicing persists much longer than bursts) as well as by appeal to information in the other spectral bands: voicing provides a power spectrum that decays with frequency approximately as 1/frequency^2, whereas most other speech sounds have flatter spectra.

V. Results and Discussion

Our small study with sixteen test cases, as seen in Table 3, resulted in a 25% error rate.

Table 3: Results of the CM (closest-match) comparison. Boldfaced values represent correct identification of speaker state. *The listed states had nearly equally small distances.

	neutral	medst	hist	scream
m1	**neutral**	neutral	neutral	**scream**
m2	**neutral**	**medst**	neutral	**scream**
m3	**neutral**	hist* neutral	**hist**	**scream**
m4	**neutral**	**medst**	**hist**	**scream**

To test the stability of the results, we performed a Principal Components Analysis (PCA, or, equivalently, singular value decomposition, SVD [9]). This permitted us to discard several of the principal components (PCs) that described only noise-level variations in the data. Retaining eight of the original 16 PCs, accounting for 95% of the variance, produced only small variations in the results, and no overall degradation in accuracy.

Table 4: Results of the PCA/CM comparison. Boldfaced values represent correct identification of speaker state. *The listed states had nearly equally small distances.

	neutral	medst	hist	scream
m1	**neutral**	neutral	neutral **hist***	**scream**
m2	**neutral**	**medst**	neutral	**scream**
m3	**neutral**	hist* neutral	**hist**	**scream**
m4	**neutral** hist*	**medst**	**hist**	**scream**

Inspection of the Tables reveals that the classification has no errors for the neutral or scream states. Furthermore, most errors occurring in the other states are manifest as neutral, that is, the closest-match algorithm selects the "conservative" interpretation that the data represent no departure from the neutral state.

VII. Conclusion

We have shown that a simple knowledge-based analysis of American English speech and some measures of F_0 can classify a speaker's emotional state among four choices moderately well. We achieve 75% accuracy when comparing new data from a speaker that is already represented among the base cases. PCA indicates that this result does not depend sensitively on small details such as noise level. We are currently investigating the performance when the speaker is not so represented.

Acknowledgements

This work was supported in part by National Science Foundation grant SGER 0206940.

References

[1] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing Emotion in Speech," *Proc. ICSLP*, 1996.

[2] H.J. Fell, L.J. Ferrier, D. Sneider, and Z. Mooraj, "EVA, An early vocalization analyzer: an empirical validity study of computer categorization," *Assets '96*, pp. 57-61, 1996.

[3] H.J. Fell, J. MacAuslan, L.J. Ferrier, K. Chenausky, "Automatic Babble Recognition for Early Detection of Speech Related Disorders," *J. Behaviour & Inf. Tech.*, **18**, no. 1, pp. 56-63, 1999.

[4] H.J. Fell, J. MacAuslan, L.J. Ferrier, S.G. Worst, and K. Chenausky, "Vocalization Age as a Clinical Tool," *Electronic Proc. ICSLP (Int. Conf. on Speech & Language Processing)*, Denver, 2002.

[5] J.H.L. Hansen, "SUSAS -Speech Under Simulated and Actual Stress*," Robust Speech Processing Lab.*, http://www.ee.duke.edu/Research/Speech/, 1997

[6] T. Johnstone, and K.R. Scherer, "The effects of emotions on voice quality*", Proc. XIVth Int. Congress of Phonetic Sci*, 1999.

[7] S. Liu, "Landmark detection of distinctive feature-based speech recognition," *J. Acc. Soc. Amer.,* **96**, 5, Part 2, p. 3227, 1994.

[8] J. MacAuslan "Speech Analysis for Fatigue Assessment", US Air Force Final Report, 2002.

[9] Press, W., S. Teukolsky, W. Vetterling, & B. Flannery. (1992). Numerical Recipes in C, 59-70. New York: Cambridge University Press.

[10] H. Quast, "Robust Machine Perception of Nonverbal Speech," *http://ergo.ucsd.edu/~holcus/Speech.html*, 2000.

[11] D. Roy & A. Pentlland, "Automatic Spoken Affect Classification and Analysis," *Pro Second Int. Conf. Automatic Face & Gesture Recognition*, pp. 363—367, 1996.

[12] K.R. Scherer, T. Johnstone, and J.Sangsue, "L'état émotionnel du locuteur: facteur négligé mais non négligeable pour la technologie de la parole," *Actes des XXIIèmes Journées d'Etudes sur la Parole,* Martigny, 1998.

[13] K.N. Stevens, S. Manuel, S. Shattuck-Hufnegel, and S. Liu, "Implementation of a model for lexical access based on features," *Proc. Int'l. Conf. Spoken Language Processing*, Banff, Alberta, **1**, 499-502, 1992.

[14] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech", *Computational Linguistics* **26**(3), pp. 339-373, 2000.