Prosody Analysis for Speaker Affect Determination

Andrew Gardner and Irfan Essa

Graphics, Visualization and Usability Center
Georgia Institute of Technology,
Atlanta, GA 30332-0280
gardner@cc.gatech.edu, irfan@cc.gatech.edu
www.cc.gatech.edu/gvu

Introduction

Speech is a complex waveform containing verbal (*e.g.* phoneme, syllable, and word) and nonverbal (*e.g.* speaker identity, emotional state, and tone) information. Both the verbal and nonverbal aspects of speech are extremely important in interpersonal communication and human-machine interaction. However, work in machine perception of speech has focused primarily on the verbal, or content-oriented, goals of speech recognition, speech compression, and speech labeling. Usage of nonverbal information has been limited to speaker identification applications. While the success of research in these areas is well documented, this success is fundamentally limited by the effect of nonverbal information on the speech waveform. The extra-linguistic aspect of speech is considered a source of variability that theoretically can be minimized with an appropriate preprocessing technique; determination of such robust techniques is however, far from trivial.

It is widely believed in the speech processing community that the nonverbal component of speech contains higher-level information that provides cues for auditory scene analysis, speech understanding, and the determination of a speaker's psychological state or conversational tone. We believe that the identification of such nonverbal cues can improve the performance of classic speech processing tasks and will be necessary for the realization of natural, robust human-computer speech interfaces. In this paper we seek to address the problem of how to systematically analyze the nonverbal aspect of the speech waveform to determine speaker affect, specifically by analyzing the pitch contour.

Methodology and Experimentation

There are many features available that may be useful for classifying speaker affect: pitch statistics, short-time energy, long-term power spectrum of an utterance, speaking rate, phoneme and silence durations, formant ratios, and even the shape of the glottal waveform [3, 4, 5, 8, 9]. Studies show, that prosody is the primary indicator of a speaker's emotional state [1, 7, 12]. We have chosen to analyze prosody as an indicator of affect since it has a well-defined and easily measureable acoustical correlate -- the pitch contour.

In order to validate the use prosody as an indicator for affect and to experiment with real speech, we need to address two problems:

- First, and perhaps most difficult, is the task of obtaining a speech corpus containing utterances that are truly representative of an affect.
- Second, what exactly are the useful features of the pitch contour in classifying affect? Especially as many factors influence the prosodic structure of an utterance and only one of these is speaker's emotional state [6, 7, 9].

In our ongoing research effort, we have considered the two above-mentioned issues separately. The first task addressed was the development of an appropriate speech corpus. In our initial experiments we asked a group of 16 students to recite phrases like "Here's looking at you, kid!", "Are you crazy?", "Are you looking at me?", *etc*., with the affects of *gladness, sadness, anger, disgust, fear, surprise,* or *neutral,* that were appropriate for each statement . The students repeated their attempts until they felt that their utterance was representative of the desired affect.

Undertaking human subject studies with this data, we found, however, that the different affects present in a specific phrase by a single speaker were marginally distinguishable to listeners not present during recording. In fact, many speakers seemed to impart little or no affect to any of the phrases. To avoid using speech that did not correctly represent affect for training, we considered using motion pictures and television broadcasts as alternate sources of utterances. We decided that day-time talk

shows were a suitable source of data for the following reasons; (a) many affect classes are clearly, even grossly, represented, (b) the speech quality is good with little background noise, and (c) much more data is available due to the quantity of daily talk shows. We have recorded several hours of programs and utterances representative of each of the affect classes have been digitized (16 kHz sampling rate, 16 bits-per-sample) and stored.

The pitch contour is extracted from the speech waveform using the method described in [12]. This is a very simple, less robust algorithm with many limitations. To improve our analysis we are implementing two different pitch detectors. One is based on the cochlear correlogram, as in [11], and the other uses PSOLA/SOLA methods. These methods will automate our pitch tracking and allow for more robust analysis. We have also identified three key features of the pitch contour that might be useful for classification: phoneme duration statistics, silence duration statistics, and pitch value statistics. In our initial studies we have chosen to examine only the pitch value statistics. We have been able to observe significant differences between different affect classes using only the mean and variance of the pitch value; this agrees with work from [13].

Conclusions and Future Work

Our next step is to perform principal component analysis on the pitch statistic feature vectors; this will require more data for our speech corpus. In future work we plan to label phonemes, words, phrases, and silences so that we may investigate their statistics for classification and clustering. We are also implementing a multiresolution decomposition method for the pitch contour using filterbanks. This is based on the hypothesis that there is an underlying macroprosodic component of the pitch contour that conveys affect [10]. Lastly, pitch contours may be "completed" across unvoiced regions of speech using interpolating splines to obtain a smooth time-frequency pitch function.

In this paper we present an approach for investigating the prosody of speech for the determination of speaker state or affect. We are also working on developing a novel speech corpus that overcomes the problem of obtaining speech truly representative of an affect class. We maintain that speaker-state classification will prove valuable to researchers investigating the production of natural-sounding synthetic speech, developing perceptual human-computer interfaces, and improving traditional speech processing algorithms.

Further details on this work are available from `www.cc.gatech.edu/~irfan/publications/pui.97/.`

References

1. R. Collier, "A comment of the prediction of prosody," in *Talking Machines: Theories, Models, and Designs*. G. Bailly, C. Benoit, and T.R. Sawallis (editors). Elsevier Science Publishers, Amsterdam: 1992.
2. W. Hess, *Pitch Determination of Speech Signals: Algorithms and Devices* Springer-Verlag, Berlin: 1983.
3. H. Kuwabara and Y. Sagisaka, "Acoustic characterstics of speaker individuality: Control and conversion", *Speech Communication*, v. 16, pp. 165-173, 1995.
4. A. Protopapas and P. Lieberman, "Fundamental frequency of phonation and perceived emotional stress", *Journal of Acoustical Society of America*, v. 101, n. 4, pp. 2267-77, 1997.
5. A. Monaghan and D. Ladd, "Manipulating synthetic intonation for speaker characterization", *ICASSP*, pp. 453-456, v. 1, 1991.
6. A. Ichikawa and S. Sato, "Some prosodical characteristics in spontaneous spoken dialogue", *International Conference on Spoken Language Processing*, v. 1, pp. 147-150, 1994.
7. D. Hirst, "Prediction of prosody: An overview", in *Talking Machines: Theories, Models, and Designs*. G. Bailly, C. Benoit, and T.R. Sawallis (editors). Elsevier Science Publishers, Amsterdam: 1992.
8. K. Cummings and M. Clements, "Analysis of the glottal excitation of emotionally styled and stressed speech", *Journal of the Acoustical Society of America*, v. 98, n. 1, pp. 88-98, 1995.
9. D. Roy and A. Pentland, "Automatic spoken affect classification and analysis", *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, pp. 363-367, 1996.
10. D. Hermes, "Pitch Analysis", in Visual Representations of Speech Signals, M. Cooker, S. Beet, and M. Crawford (editors). Wiley and Sons, New York: 1993.
11. M. Slaney and R. Lyon, "On the importance of time-- a temporal representation of sound", in Visual Representations of Speech Signals, M. Cooker, S. Beet, and M. Crawford (editors). Wiley and Sons, New York: 1993.
12. L. Rabiner and R. Shafer, *Digital Processing of Speech Signals*, Wiley and Sons, New York: 1978.
13. W. Williams, and Stevens, K. N. "Emotions and speech: Some acoustical correlates." *Journal of the Acoustical Society of America*, v. 52, n. 4, pp. 1238 - 1250, 1972.
14. J. Cahn, “Generation of Affect in Synthesized Speech”, In *Proceedings of AVIOS 89*, pp. 251-256, 1989.