Realizing Affect in Speech Classification in Real-Time

Carson Reynolds and **Masatoshi Ishikawa**
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
{carson,ishikawa}@k2.t.u-tokyo.ac.jp

Hiroshi Tsujino
Honda Research Institute Japan Co. Ltd.
8-1 Honcho, Wako-shi, Saitama, 351-0114, Japan
tsujino@jp.honda-ri.com

Abstract

Robots can recognize pitch, perceived loudness, or simple categories such as positive or negative attitude in real-time from speech. Existing research has demonstrated specialized systems able to perform these operations. Additionally, off-line classification has shown that affect can be classified at rates that are better than random although worse than human performance. Taken together, this previous research indicates the possibility of real-time classification of a variety of affective states. We are exploring this problem with the eventual goal of designing an field-programmable gate array (FPGA) based system that will rapidly process relevant features in parallel.

Harsh Words and Robotics

It would be useful and amusing to make robots capable of responding to emotional information in human speech. Indeed this information could be used to train robots using techniques like reinforcement learning. The muttering of profanities to un-comprehending computers or automobiles seems to indicate an innate desire for these machines to respond.

This paper examines the possibility of designing a system that would allow rapid response to the affective content of speech. As a starting point, existing systems for real-time processing of different aspects of speech are reviewed. Proceeding from this we will describe some work-in-progress to design a FPGA that processes information derived from speech in parallel.

Robotics and Emotions

In addition to the fiction of Kapek and Asimov, a variety of literature has discussed the possibility of robots having emotions (Sloman & Croucher 1981), responding to emotions (Picard 1997), and displaying emotions (Breazeal & Scassellati 1999). There are also a number of actual robots such as Kismet and Sparky (Scheeff *et al.* 2000), which seek to display emotions.

With robotics there has been a focus on "output" of models of emotional states while "input" still remains impoverished. This is likely because systems to recognize emotional states remain an area of active research.

Speech and Emotion

Emotion in speech can be conveyed in a variety of ways. Prosodic information such as pitch, loudness, rhythm, and stress can convey information about emotional states. Paralinguistic information such as attitude, intention, and physiological state may also provide information related to emotion. A plurality of definitions and an ongoing debate surround the exact usage of the terms prosody and paralinguistic (Schotz 2003). Emotion may also be conveyed in the semantic information of spoken speech. However, one conundrum facing researchers of emotions is that there are no generally agreed upon models for emotions (Cowie *et al.* 2001).

Initial approaches to the problem of recognizing emotional content of speech have often used statistical pattern recognition techniques. For instance in "Recognizing Emotion in Speech" describes a system that classifies utterances using prosodic features and pattern recognition techniques (Dellaert, Polzin, & Waibel 1996). Recordings from speakers told to convey one of four emotions (happiness, sadness, anger, and fear) were identified with 82% accuracy. More recently work has been performed to rank various features in terms of their discriminative performance in classifying emotion in speech (Fernandez & Picard 2005). This work has begun to approach human classification performance of speech in emotion. However, much of this pattern recognition is performed in an off-line manner.

Real-time Processing of Prosodic Features

A variety of projects show that different prosodic features, however can be computed in real time. The following sections review a few of these projects, as grouped by prosodic feature.

Loudness

Transducing loudness in real-time is relatively straightforward, but making use of perceptual models for loudness provides introduces some complexity. Using these models allows a more realistic approximation of how the human ear attenuates and perceives raw sound. One example

of a real-time system is provided by Tuomi and Zacharov who describe an implementation of the Equivalent Rectangular Bandwidth (ERB) loudness model (Tuomi & Zacharov 2000).

Robot audition provides special problems stemming from noise due to motor activity. Nakadai et al. describe real-time processing of loudness information while removing motor noise with the aim of locating sound sources (Nakadai *et al.* 2001).

Pitch

Many prosodic features are derived from pitch and the various algorithms by which it is extracted from the speech signal. Dubnowski et al. describe a hardware pitch detector that operates at 10 kHz. Using a clipping and a simplified autocorrelation algorithm, they show a working hardware implementation (even provding schematics) (Dubnowski, Schafer, & Rabiner 1976).

Attitude

Making use of prosodic information such as the fundamental frequency (F0) contours and phoneme alignment, Fujie et al. describe a system that classifies negative or positive attitude in real-time (Fujie *et al.* 2005). However, the system is not hardware-based instead using ESPS C library for pitch tracking.

ChAff

We have now seen a set of systems that use prosodic features to classify emotional states. Additionally, we have seen a variety of systems which compute some of these prosodic features in real-time. Based on this we argue that is now possible to build a system which classifies the emotional state of speech in real time.

The Chip for real-time classification of Affect (ChAff) project is seeking to take many of these existing real-time features and parallelize them. The approach of the project is to build simulations making use of Mathwork's Simulink environment. After the simulations are validated we plan to translate the underlying logic into a register transfer level (RTL) description suitable for synthesis on an field programmable gate array (FPGA) based embedded system.

At the moment the project is in its initial stages. Thus far we have implemented simple simulations. One real-time simulation computes a root-mean-square (RMS) normalized version of an input signal and feeds it into a filterbank whose center frequencies and bandwidths are defined by the Bark-scale (Smith & Abel 1995). These signals are in turn subjected to short-time analysis and thresholded to provide an average of energy in particular frequency bands.

The project's next steps are to implement a full-realized perceptual loudness model such as Zwicker's (Zwicker & Fastl 1999). Following this, an implementation of an autocorrelating / clipping pitch detector will follow.

At the present, we are making use of existing corpora of speech which have been labeled with categories. However, as the project progresses it is our hope to collect a new corpus that better reflects the goal of realizing affect in speech classification in real-time in the domain of robotics. We are considering developing an on-line "guess my emotion" game that will provide both speech samples and a variety of labels.

References

Breazeal, C., and Scassellati, B. 1999. How to build robots that make friends and influence people. In *Intelligent Robots and Systems, 1999. IROS '99. Proceedings. 1999 IEEE/RSJ International Conference on*, volume 2, 858–863 vol.2.

Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; and Taylor, J. G. 2001. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE* 18(1):32–80.

Dellaert, F.; Polzin, T.; and Waibel, A. 1996. Recognizing emotions in speech. In *Proc. ICSLP '96*, volume 3, 1970–1973.

Dubnowski, J.; Schafer, R.; and Rabiner, L. 1976. Real-time digital hardware pitch detector. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24(1):2–8.

Fernandez, R., and Picard, R. W. 2005. Classical and novel discriminant features for affect recognition from speech. In *nterspeech 2005 - Eurospeech 9th European Conference on Speech Communication and Technology*.

Fujie, S.; Kobayashi, T.; Yagi, D.; and Kikuchi, H. 2005. Prosody based attitude recognition with feature selection and its application to spoken dialog system as para-linguistic information. In *INTERSPEECH-2004*, 2841–2844.

Nakadai, K.; Hidai, K. I.; Mizoguchi, H.; Okuno, H. G.; and Kitano, H. 2001. Real-time auditory and visual multiple-object tracking for humanoids. In *IJCAI*, 1425–1436.

Picard, R. W. 1997. *Affective Computing*. The MIT Press.

Scheeff, M.; Pinto, J.; Rahardja, K.; Snibbe, S.; and Tow, R. 2000. Experiences with sparky: A social robot. In *Proceedings of the Workshop on Interactive Robot Entertainment*.

Schotz, S. 2003. Prosody in relation to paralinguistic phonetics - earlier and recent definitions, distinctions and discussions.

Sloman, A., and Croucher, M. 1981. Why robots will have emotions. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81), Vancouver, BC, Canada, August 1981*. William Kaufmann.

Smith, J. O., and Abel, J. S. 1995. The bark bilinear transform. In *Applications of Signal Processing to Audio and Acoustics, 1995., IEEE ASSP Workshop on*, 202–205.

Tuomi, O., and Zacharov, N. 2000. A real-time binaural loudness meter. In *Presented at the 139th meeting of the Acoustical Society of America*.

Zwicker, E., and Fastl, H. 1999. *Psychoacoustics : Facts and Models*. Springer Series in Information Sciences. Springer.