Feature Combination for Better Differentiating Anger from Neutral in Mandarin Emotional Speech

Tsang-Long Pao[1], Yu-Te Chen[1], Jun-Heng Yeh[1], Yun-Maw Cheng[1],
and Charles S. Chien[2]

[1] Department of Computer Science and Engineering, Tatung University
[2] School of Management and Development, Feng Chia University
{tlpao,kevin}@ttu.edu.tw, {d8906005,d9306002}@ms2.ttu.edu.tw,
scchien@fcu.edu.tw

Extended Abstract

Just as written language is a sequence of elementary alphabet, speech is a sequence of elementary acoustic symbols. Speech signals convey more than spoken words. The additional information conveyed in speech includes gender information, age, accent, speaker's identity, health, prosody and emotion [1].

Affective computing, which is currently a very attractive research topic, aims at the automatic recognition and synthesis of emotions in speech, facial expressions, or any other biological signals [2]. Recently acoustic investigation of emotions expressed in speech has gained increased attention partly due to the potential value of emotion recognition for spoken dialogue management [3-4]. Imagine for example a call-center system that can detect complain or anger due to unsatisfied services about user's requests, could deal it smoothly by transferring the user to human operator. However, in order to reach such a level of performance we need to extract a reliable acoustic feature set that is largely immune to inter- and intra-speaker variability in emotion expression. Within the field of affective computing, this paper addresses how far we can go to use feature combination (FC) to concatenate different features to improve the accuracy of differentiating anger from neutral emotion in Mandarin speech.

We selected the pitch, log energy, formants, linear predictive coefficients (LPC), linear prediction cepstral coefficients (LPCC), mel-frequency cepstral coefficients (MFCC), log frequency power coefficients (LFPC), perceptual linear prediction (PLP), relative spectral PLP (Rasta-PLP), jitter and shimmer as the base features. We also added velocity and acceleration information for pitch and MFCCs, respectively, to take the speaking rate into account and model the dynamics of the corresponding temporal change of pitch and spectrum.

Feature combination is a well-known technique [5]. During the feature extraction of speech recognition system, we typically find that each feature type has particular circumstances in which it excels, and this has motivated our investigations for combining separate feature streams into a single emotional speech recognition system. However, it is forbiddingly time consuming to perform exhaustive search for the subset of features that give best classification. Due to the highly redundant information in the concatenated feature vector, a forward feature selection (FFS) or backward feature selection (BFS) should be carried out to extract only the most representative features, thereby orthogonalizing the feature vector and reducing its dimensionality.

A. Paiva, R. Prada, and R.W. Picard (Eds.): ACII 2007, LNCS 4738, pp. 741–742, 2007.

The used corpus in this paper is constructed by MIR lab at National Tsing Hua University in Taiwan. They invite one female to portray two emotions, including anger and neutral. Finally, they obtained 2000 utterances that were recorded in 16-bit PCM with a sampling frequency of 16k Hz including 1000 angry and 1000 neutral utterances.

The Mandarin emotion recognition system was implemented using "MATLAB" software run under a desktop PC platform. The correct recognition rate was evaluated using leave-one-out (LOO) cross-validation which is K-fold cross validation taken to its logical extreme, with K equal to the number of data points.

The task of the classifier component proper of a full system is to use the feature vector provided by the feature extractor to assign the object to a category. To recognize emotions in speech we tried the following approaches: linear discriminant analysis (LDA), support vector machine (SVM), and back-propagation neutral network (BPNN), as they had been applied successfully in previous researches and achieved high accuracy [6-7].

The experimental results show that Rasta-PLP is the most important feature in LDA and BPNN classifiers and accMFCC is the most important one in SVM classifier. The velMFCC or accMFCC that model the dynamics of the corresponding temporal change spectrum can achieve better accuracy than MFCC. Jitter is the most irrelevant feature among these classifiers. Contrary to [6], the pitch and energy were not shown to play a major role in this paper. By Adopting feature selection and combination, the recognition rates can improve 2.1%, 7.8% and 20.6%, respectively. The best accuracy of differentiating anger from neutral can be achieved 99.10% by using accMFCC, LFPC, LPC, PLP, and RASTA-PLP feature streams and the support vector machine classifier. Although BPNN achieved the lowest recognition rate among all classifiers, it benefited most by feature combination in our system.

References

1. Rabiner, L.R., Juang, B.H.: Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ (1993)
2. Picard, R.W.: Affective Computing. MIT Press, Redmond, Washington (1997)
3. Lee, C.M., Narayanan, S.: Towards detecting emotion in spoken dialogs. IEEE Trans. on Speech & Audio Processing (in press)
4. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S.: Emotion Recognition in Human-Computer Interactions. IEEE Sig. Proc. Mag. 18, 32–80 (2001)
5. Ellis, D.: Stream combination before and/or after the acoustic model. In: Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (2000)
6. Kwon, O.W., Chan, K., Hao, J., Lee, T.W.: Emotion Recognition by Speech Signals. In: Proceedings of EUROSPEECH, pp. 125–128 (2003)
7. Bhatti, M.W., Wang, Y., Guan, L.: A Neural Network Approach for Human Emotion Recognition in Speech. In: Proceedings of the 2004 International Symposium, vol. 2, pp. 1184–1811 (2004)