Affect-Insensitive Speaker Recognition by Feature Variety Training

Dongdong Li and Yingchun Yang*

Department of Computer Science and Technology,
Zhejiang University, Hangzhou, P.R. China, 310027
{lidd,yyc}@zju.edu.cn

A great deal of inner variabilities such as emotion and stress are largely missing from traditional speaker recognition system. The direct result is that the recognition system is easily disturbed when the enrollment and the authentication are made under different emotional state. Reynolds [1] proposed a new normalization technique called feature mapping. This technique achieved big successes in channel robust speaker verification. We extend the mapping idea to develop a feature variety training approach for affective-insensitive speaker recognition.

The feature variety training algorithm could be implemented by two processes: model parameters shift and feature transformation. First, the transformations are learned by examining how model parameters shift after MAP adaptation which builds the emotional depended models from the background one. Then different emotional types of generated features are obtained for speaker model building.

Model parameters shift: An emotional independent background model (UBM) is trained using all available emotional data including the neutral speech. Next, emotional specific models (EDM) are derived by MAP adaptation with emotional specific data that is also used to generate the UBM. All models are derived with a common background, which lead to a correspondence between Gaussian components in the models. Let $E_i = [e_{i1}, e_{i2}, ..., e_{iN}]$ be the i^{th} type of specific affective feature used to construct EDMi, where N is the frame number of training data E_i. The top-1 decode Gaussian j_i for the i^{th} type of emotional speech is determined by

$$j_i = \arg\max_{1\le k\le M} \sum_{n=1}^{N} l_k^{EDM_i}(e_{in}) = \arg\max_{1\le k\le M} \sum_{n=1}^{N} \omega_k^{EDM_i} p_k^{EDM_i}(e_{in}) \quad (1)$$

where $p_k^{EDM_i}(e_{in}) = N(\mu_k^{EDM_i}, \sigma_k^{EDM_i})$ is the k^{th} mixture component of the GMM EDM_i and M is the orders of the Gaussian mixture model.

Feature transformation: Let $\mu_{j_i}^{EDM_i}$ and $\sigma_{j_i}^{EDM_i}$ be the mean and the standard deviation of the top-1 order with the maximal likelihood in the EDMi. Given enrollment speech x, the transformed feature of the i^{th} type of emotional space y_i, is then given by

$$y_i = FT_i(x) = (x - \mu_{j_i}^{UBM}) * \sigma_{j_i}^{EDM_i} / \sigma_{j_i}^{UBM} + \mu_{j_i}^{EDM_i} \quad (2)$$

* Corresponding author.

A. Paiva, R. Prada, and R.W. Picard (Eds.): ACII 2007, LNCS 4738, pp. 743–744, 2007.

where $\mu_{j_i}^{UBM}$ and $\sigma_{j_i}^{UBM}$ is the mean and the standard deviation of the corresponding order j in the UBM. The i^{th} type of emotional speaker model is trained with the generated feature y_i.

The enrollment speech is used to generate the target emotion speech of all types with the feature transformation function. The speaker models (SM) of different emotional types are established. For the authentification process, the log likelihood of the test (identified or verified) utterance is computed against the most likely type of emotional speaker model (SM). The maximum likelihood rule is applied.

The experiments are conducted with the Mandarin Affective Speech Corpus (MASC) [2]. Two protocols are defined to evaluate the performance of speaker authentication. In Protocol I, both the SM training and the pre-build model (like the UBM and the EDM) training data are drawn from the same speaker sets. For Protocol II, the learned model parameters are totally separated from the test set of the system. In each strategy, the Hamming windows size is 32 ms. The feature vector is composed by 16 dimensional mel-cepstral and its Delta. The silence and unvoiced segments are discarded based on an energy threshold. The models are 1024 order Gaussian Mixture Models. ALIZE [3] is used as the interface in our source code. Table 1 reports the identification rate (IR) of the speaker authentication system observed over different types of affective speech compared with the GMM-UBM in two protocols.

Table 1. The identification rate of the system(%)

IR	GMM-UBM		Feature Variety Training	
	Protocol I	Protocol II	Protocol I	Protocol II
Neutral	0.9243	0.9483	0.9450	0.9712
Anger	0.3004	0.3738	0.3195	0.3933
Elation	0.3265	0.4067	0.3357	0.4300
Panic	0.2599	0.3567	0.2710	0.3733
Sadness	0.5114	0.6029	0.5331	0.6238

It is significant to note that similar profits of performance in protocol Ⅰ and protocol Ⅱ are achieved compared the baseline and feature variety training strategy, while in the second protocol the training and test data of the speakers is totally separated from the UBM and EDM training data. The encouraging result demonstrates that feature variety training could be learned in advance from the prior affective speech irrelevant to the registered users by the proposed technique.

Acknowledgments. This work is supported by NSFC_60525202/60533040, 863 Program_2006AA01Z136, PCSIRT0652, ZPNSF_Y106705, NCET-04-0545.

References

1. Reynolds, D.A.: Channel robust speaker verification via feature mapping. ICASSP'03 2, 53–56 (2003)
2. Wu, T., Yang, Y.C., Wu, Z.H., Li, D.D., MASC,: A Speech Corpus in Mandarin for Emotion Analysis and Affective Speaker Recognition, The IEEE Odyssey, pp. 1–59 (2006)
3. Bonastre, J.F., Wils, F., Meignier, S.: ALIZE, a free toolkit for speaker recognition, IEEE International Conference on Acoustics, Speech, and Signal Processing. In (ICASSP '05), March 18-23, vol. 1, pp. 737–740 (2005)