How Does Noise Impact Speech-based Emotion Classification?

Na Yang
University of Rochester
Rochester, NY 14627, USA
nayang@rochester.edu

Ilker Demirkol
Universitat Politecnica de Catalunya
Barcelona, Catalunya, Spain
ilker.demirkol@entel.upc.edu

Jianbo Yuan
University of Rochester
Rochester, NY 14627, USA
jyuan10@ece.rochester.edu

Wendi Heinzelman
University of Rochester
Rochester, NY 14627, USA
wendi.heinzelman@rochester.edu

Yun Zhou
University of Rochester
Rochester, NY 14627, USA
yzhou43@ece.rochester.edu

Melissa Sturge-Apple
University of Rochester
Rochester, NY 14627, USA
melissa.sturge-apple@rochester.edu

CHI'14, April 26–May 1, 2014, Toronto, Canada.

DOI string from ACM form confirmation

Abstract
As an essential approach to understanding human interactions, emotion classification is a vital component in the design of human-computer interaction (HCI) systems. Speech contains rich information about emotion, but the impact of noise on the classification performance is still not well studied, especially for applications used in noisy mobile environments. For an emotion classification system using support vector machine with a threshold-based fusion mechanism, we study the impact of noisy speech data on the performance of emotion classification for a standard emotion database.

Author Keywords
Emotion classification, support vector machine, thresholding fusion, noisy speech

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous.

Introduction
Speech contains rich information for effectively conveying emotions in the communications between humans, and this has motivated researchers to explore the area of emotion classification based on speech [6] [3] and the broader HCI domain [5]. Mining useful emotion

information solely from prosodic features is still a challenging task, and a thorough analysis of the emotion classification accuracy under real scenarios is necessary, such as where modalities are captured in noisy environments.

The emotion classification system used in this paper extracts speech features, and the widely employed Support Vector Machine (SVM) learner is used for One-Against-All (OAA) classification for each emotion. The confidence levels from individual OAA classifiers are combined by means of a thresholding fusion mechanism to improve the classification performance. We compare the six-emotion classification performance for an original database with clean speech and speech data with added babble or white noise.

Mobile applications

Speech-based emotion classification can be an entry point for elaborate context-aware systems for the future mobile market. For example, smartphones may be customized to automatically choose songs or background colors based on the user's current emotion. Voice blogging on social voice platforms, such as Bubbly and Twitterfone, also enables sociologists to study emotion states of the mass population from social media. In the healthcare field, speech-based emotion sensing technologies have been implemented on mobile devices for behavioral studies [6] or patient monitoring [8].

However, for these types of applications, the various ambient noise captured by mobile devices may strongly influence the accuracy of the speech features detection, and therefore the emotion classification performance could be greatly influenced.

Emotion classification system

In this section, a multiclass SVM with thresholding fusion for speech-based emotion classification is presented.

Speech features evaluated

In order to maintain a low computational complexity of the system, we only choose the basic and commonly-used speech features as the attributes for emotion classification: fundamental frequency (F_0), energy, frequency and bandwidth of the first four formants, Mel-scale Frequency Cepstrum Coefficients (MFCCs), speaking rate, the difference of F_0 and the difference of energy between neighboring frames. We divide each speech utterance into 60 ms segments with 10 ms time shifts, and only analyze the speech features for voiced segments. We use the noise-resilient BaNa F_0 detection algorithm [1] to extract the F_0 values.

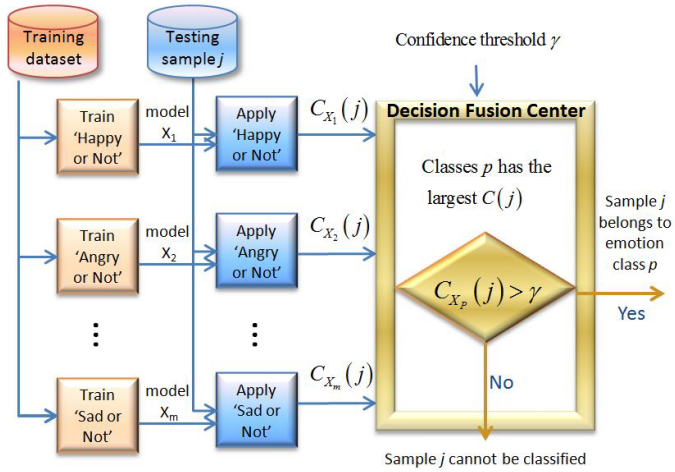

Figure 1: The emotion classification system using OAA SVM with thresholding fusion.

For each speech utterance, we calculate five statistics: the mean, maximum, minimum, range, and standard deviation for each feature vector except speaking rate. The z-score speaker normalization scheme is applied to reduce the inter-speaker variability and increase the emotion classification accuracy.

Multiclass emotion classification with thresholding fusion We implement the one-against-all (OAA) approach for emotion classification using multiclass SVM and the RBF kernel to deal with the non-linear relationship between the class labels and the features. The SMOTE method [2] is used to upsample the uneven dataset for binary classification for individual OAA classifiers.

Figure 1 illustrates the OAA SVM classification system with the thresholding fusion mechanism that we use from [7]. In the testing phase, the confidence measures $C_{X_i}(j)$ from all OAA classifiers are sent to the fusion center, where model X_p yields the highest confidence measure for utterance j. We use the thresholding fusion mechanism proposed in [7] to compare this largest confidence value $C_{X_p}(j)$ against a user-controlled confidence threshold γ to decide whether to reject the sample as unclassified.

Evaluation

The widely-used LDC dataset [4] is chosen for performance evaluation, which includes 727 utterances recorded by 3 professional actors and 4 actresses reading semantically neutral-meaning utterances. Six emotions are selected in our emotion classification study: disgust, happiness, sadness, anger, fear and neutral.

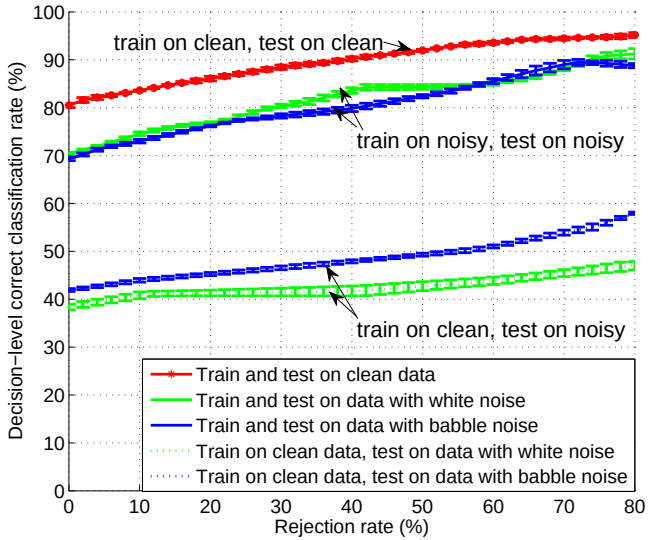

Figure 2: Decision-level correct classification rate vs. rejection rate for cross-validation tests on clean and noisy LDC data at 5 dB SNR.

The decision-level correct emotion classification rates for cross-validation tests on clean LDC data is presented in Fig. 2. The error bars indicate the performance variations among 5 times SMOTE upsampling on the uneven training datasets. When no data is rejected, the classification rate is 80%, which is much better than a random guess result, i.e., 1/6=16.7%. This number can be increased to 95% when 80% of the data is rejected. Therefore, using the thresholding fusion method can provide a more reliable emotion classification at the expense of leaving some data unclassified.

Since white noise and babble noise are two common types

of noise, we add these two types of noise to the LDC speech signals to generate a noisy dataset. A moderate noise level, i.e., noisy data at 5 dB SNR, is used for testing. We can see from Fig. 2 that for emotion classification on noisy data, the correct classification rate for training the system using noisy data is around 75% higher than the correct classification rate for training using clean data. Though speaker normalization could help to combat the overall increase in energy for the noisy data, it does not help with features in the frequency domain. When trained with noisy data, the system can, on the other hand, learn the spectral features for noisy speech. Therefore, the classification rate does not drop too much for training and testing on noisy data.

Conclusions

For emotion classification in real scenarios, noise is a factor that inevitably needs to be considered for performance evaluations. We discuss several noisy scenarios that may require speech-based emotion classification. Experimental results show the impact of noise on the emotion classification performance.

For interaction designers and HCI participants designing interactive systems using users' emotion from speech, the noise effect should be taken into consideration as an important factor. To more effectively classify emotion on noisy data, the system should be trained using noisy data instead of clean data. To reduce the influence of noise on system reliability, we can adapt the system to different noise levels. For example, we can choose to increase the confidence score threshold used in the SVM thresholding fusion for very noisy scenarios, and only classify emotions when the confidence score is relatively high. Additionally, we can sample the user's speech multiple times within a short period of time, and derive the user's emotion based on the majority classified emotion on these samples.

References

[1] Ba, H., Yang, N., Demirkol, I., and Heinzelman, W. BaNa: A hybrid approach for noise resilient pitch detection. In *IEEE Statistical Signal Processing Workshop* (2012), 369 –372.

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research 16* (2002), 321–357.

[3] Hoque, M., Yeasin, M., and Louwerse, M. Robust recognition of emotion from speech. In *Intelligent Virtual Agents*, vol. 4133 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2006, 42–53.

[4] Liberman, M., Davis, K., Grossman, M., Martey, N., and Bell, J. Emotional prosody speech and transcripts. *Linguistic Data Consortium* (2002).

[5] Munteanu, C., Jones, M., Oviatt, S., Brewster, S., Penn, G., Whittaker, S., Rajput, N., and Nanavati, A. We need to talk: HCI and the delicate topic of spoken language interaction. In *CHI '13 Extended Abstracts* (2013), 2459–2464.

[6] Rachuri, K. K., Musolesi, M., Mascolo, C., Rentfrow, P. J., Longworth, C., and Aucinas, A. EmotionSense: A mobile phones based adaptive platform for experimental social psychology research. In *Proc. of UbiCom* (2010), 281–290.

[7] Vapnik, V. N. *Statistical Learning Theory*. Wiley, 1998.

[8] Yang, Y., Fairbairn, C., and Cohn, J. F. Detecting depression severity from vocal prosody. *Affective Computing, IEEE Tran. on 4*, 2 (2013), 142–150.