Fisher Kernels on Phase-based Features for Speech Emotion Recognition

Jun Deng[1], Xinzhou Xu[2], Zixing Zhang[1], Sascha Frühholz[3], Didier Grandjean[3], and Björn Schuller[1,3,4]

Abstract The involvement of affect information in a spoken dialogue system can increase the user-friendliness and provide a more natural way for the interaction experience. This can be reached by speech emotion recognition, where the features are usually dominated by the spectral amplitude information while they ignore the use of the phase spectrum. In this paper, we propose to use phase-based features to build up such an emotion recognition system. To exploit these features, we employ Fisher kernels. The according technique encodes the phase-based features by their deviation from a generative Gaussian mixture model. The resulting representation is fed to train a classification model with a linear kernel classifier. Experimental results on the GeWEC database including 'normal' and whispered phonation demonstrate the effectiveness of our method.

1 Introduction

For a spoken dialogue systems, a recent trend is to consider the integration of emotion recognition in order to increase the user-friendliness and provide a more natural interaction experience [3, 1, 20, 6, 28, 5]. In fact, this may be particularly relevant for systems that accept whispered speech as input given the social and emotional implications of whispering. At present, acoustic features used for speech emotion recognition are dominated by the conventional Fourier transformation magnitude part of a signal, such as in Mel-frequency cepstral coefficients (MFCCs) [7, 21, 2, 4]. In general, the phase-based representation of the signal has been neglected mainly because

[1] Chair of Complex & Intelligent Systems, University of Passau, Passau, Germany, e-mail: jun.deng@uni-passau.de
[2] Machine Intelligence & Signal Processing group, MMK, Technische Universität München, Munich, Germany
[3] Swiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland
[4] Department of Computing, Imperial College London, London, UK

of the difficulties in phase wrapping [15, 31]. In spite of this, the phase spectrum is capable of summarising the signal. Recent work has proved the effectiveness of using phase spectrum in different speech audio processing applications, including speech recognition [17, 12], source separation [16], and speaker recognition [13]. However, there exists little research, which applies phase-based features for speech emotion recognition. Recently, the phase distortion, which is the derivative of the relative phase shift, has been investigated for emotional valence recognition [27]. In this short paper, the key objective is to demonstrate the usefulness of the phased-based features for speech emotion recognition. In particular, this paper investigates whether the *modified group delay feature* is capable of improving the performance of an emotion recogniser because it has not yet been applied for speech emotion recognition. Besides, we propose to use *Fisher kernels* to encode the varied length series of the modified group delay features into a fixed length Fisher vector. The Fisher kernel is a powerful framework, which enjoys the benefits of generative and discriminative approaches to pattern classification [14]. Eventually, a linear kernel support vector machine (SVM) is adopted to train the emotion recognition model with the resulting Fisher vectors.

2 Methods

2.1 Modified Group Delay Feature

Given a discrete time signal $x(n)$, the group delay feature is written as follows

$$\tau_g(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2}, \tag{1}$$

where the angular frequency ω is limited in $[0, 2\pi]$, n is an integer, $|X(\omega)|$ is the magnitude of the Fourier transforms of $x(n)$, $Y(\omega)$ is the Fourier transforms of the signal $y = nx(n)$, and the subscripts R and I indicate real and imaginary parts, respectively. Although the group delay feature is discriminative and additive, it is ill behaved if the zeros of the system transfer function are close to the unit circle [17]. To address this issue, a modification of the group delay function is proposed in [17]. The modified feature is computed as

$$\tau_m(\omega) = \frac{\tau_p(\omega)}{|\tau_p(\omega)|}|\tau_p(\omega)|^{\alpha}, \tag{2}$$

where

$$\tau_p(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^{2\gamma}}, \tag{3}$$

and $S(\omega)$ is a smoothed form of $|X(\omega)|$. The two tuning parameters γ and α are set to 1.2 and 0.4 in this paper based on previous work [29, 30]. Similar to the com-

putation of MFCCs, the discrete cosine transform (DCT) is applied on the features computed by Equation (2) so as to perform a decorrelation. The first 12 coefficients are retained (excluding the 0^{th} coefficient).

2.2 Fisher Kernels

The Fisher kernel is popularly used in large scale image classification and image retrieval [18, 19]. The basic idea is to look at how the low level descriptors (e. g., a sequence of the phase-based features) affect the learnt generative model, which is typically a Gaussian mixture model (GMM). The effect is obtained by computing the derivative of the log-likelihood with respect to the model parameters. Formally, given a parametric generative model p_λ with parameters λ, the Fisher score function of a given example X is given by computing the first derivative of the log-likelihood function

$$\Phi(X) = \nabla_\lambda \log p_\lambda(X). \tag{4}$$

The Fisher score allows us to embed a sequence of low-level descriptors into a fixed-length vector, whose dimensionality depends on the size of the model parameters.

3 Experiments

3.1 Geneva Whispered Emotion Corpus

We employ the Geneva Whispered Emotion Corpus (GeWEC) to evaluate the effectiveness of the proposed system. The corpus provides normal and whispered paired utterances. Two male and two female professional French-speaking actors in Geneva were recruited to speak eight predefined French pseudo-words (e. g., *"belam"* and *"molen"*) with a given emotional state in both normal and whispered speech modes as in the GEMEPS-corpus that was used in the Interspeech 2013 Computational Paralinguistics Challenge [26]. Speech was expressed in four emotional states: *angry*, *fear*, *happiness*, and *neutral*. The actors were requested to express each word in all four emotional states five times. The utterances were labelled based on the state they should be expressed in, i. e., one emotion label was assigned to each utterance. As a result, GeWEC consists of 1 280 instances in total. In the experiments, cross-speech-mode evaluation is considered. That is, one speech mode of the GeWEC data is used for training while the other speech mode data is used for testing.

Table 1 UAR for four-way cross-mode emotion recognition on GeWEC: When one speech mode of GeWEC (normal speech (norm.) or whispered speech (whisp.)) is used for training, the other one is used for testing. Different feature sets are considered. Significant results (p-value < 0.05, one-sided z-test) are marked with an asterisk.

UAR [%]	IS09	IS10	IS11	IS12	IS13	GeMAPS	eGeMAPS	Proposed
Norm. (train), Whisp. (test)	35.5	39.5	40.3	33.3	36.4	34.1	41.9	$*$**50.3**
Whisp. (train), Norm. (test)	53.4	52.3	52.8	46.4	48.4	32.0	38.9	**54.8**

3.2 Experimental Setup

To generate Fisher vectors on the low level descriptors, we use a K-component GMM with diagonal covariances as the generative model p_λ. As suggested in [18], only gradients of the means and covariances are taken into account, leading to a $2 \times d \times K$ dimensional vector, where $d = 12$ for the phase-based features. The number of components for GMMs is chosen in a range $K = \{2,4,\ldots,30\}$ via cross-validation. As for the basic supervised learner in the classification step, we use linear SVMs implemented in LIBLINEAR [11].

As for the baselines, we chose to use SVMs with various state-of-the-art and publicly available feature sets, provided by the open source openSMILE toolkit [10, 9]. The feature sets used include Interspeech Challenges on Emotion in 2009 [23] (IS09), Level of Interest in 2010 [24] (IS10), Speaker States in 2011 [22] (IS11), Speaker Traits in 2012 [25] (IS12), Emotion in 2013 [26] (IS13), and the most recently proposed Geneva minimalistic acoustic parameter set (GeMAPS) and the extended Geneva minimalistic acoustic parameter set (eGeMAPS) [8]. Unweighted average recall (UAR) is used as a performance metric as in these challenges.

3.3 Results

Table 1 presents the experimental results on the GeWEC data. As can be seen from Table 1, the proposed method always performs best for the two experimental settings. For the first one, where the model is trained on 'normal' while tested on whispered speech, the phase-based features in conjunction with the Fisher kernel reaches 50.3 % UAR, which outperforms the other considered methods significantly by a large margin. As for the second setting, the proposed method achieves 54.8 % UAR, which is as competitive as the other approaches.

4 Conclusions

We focused on improving speech emotion recognition, e. g., for the embedding in spoken dialogue systems, in a challenging whispered vs non-whispered speech and vice-versa cross-mode setting. Specifically, we presented a novel framework, using phased-based features (i. e., modified group delay features) with a Fisher kernel. Cross-speech-mode experiments on the GeWEC data were conducted, demonstrating that the present framework is competitive with other modern emotion recognition models. Further evaluations of alternative types of phase-based features are to be considered. Other future work includes to incorporate different normalisation techniques such as L2 normalisation [19] into the recognition framework.

Acknowledgments

This work has been partially supported by the BMBF IKT2020-Grant under grant agreement No. 16SV7213 (EmotAsS) and the European Communitys Seventh Framework Programme through the ERC Starting Grant No. 338164 (iHEARu).

References

1. Acosta, J.C.: Using emotion to gain rapport in a spoken dialog system. In: Proc. NAACL HLT, pp. 49–54. Boulder, CO (2009)
2. Anagnostopoulos, C.N., Iliou, T., Giannoukos, I.: Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. Artificial Intelligence Review pp. 1–23 (2012)
3. Andre, E., Rehm, M., Minker, W., Bühler, D.: Endowing spoken language dialogue systems with emotional intelligence. In: Affective Dialogue Systems, pp. 178–187. Springer (2004)
4. Attabi, Y., Alam, M.J., Dumouchel, P., Kenny, P., O'Shaughnessy, D.: Multiple windowed spectral features for emotion recognition. In: Proc. ICASSP, pp. 7527–7531. IEEE, Vancouver, BC (2013)
5. Benyon, D., Gamback, B., Hansen, P., Mival, O., Webb, N.: How was your day? Evaluating a conversational companion. IEEE Transactions on Affective Computing **4**(3), 299–311 (2013)
6. Callejas, Z., Griol, D., López-Cózar, R.: Predicting user mental states in spoken dialogue systems. EURASIP Journal on Advances in Signal Processing **2011**, 6 (2011)
7. Dumouchel, P., Dehak, N., Attabi, Y., Dehak, R., Boufaden, N.: Cepstral and long-term features for emotion recognition. In: Proc. INTERSPEECH, pp. 344–347. Brighton, UK (2009)
8. Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., Truong, K.: The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing (2015). 14 pages, to appear
9. Eyben, F., Weninger, F., Groß, F., Schuller, B.: Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: Proc. MM, pp. 835–838. Barcelona, Spain (2013)
10. Eyben, F., Wöllmer, M., Schuller, B.: openSMILE – the Munich versatile and fast open-source audio feature extractor. In: Proc. MM, pp. 1459–1462. Florence, Italy (2010)

11. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. The Journal of Machine Learning Research **9**, 1871–1874 (2008)
12. Hegde, R., Murthy, H., Gadde, V.: Significance of the modified group delay feature in speech recognition. IEEE Transactions on Audio, Speech, and Language Processing **15**(1), 190–202 (2007)
13. Hernáez, I., Saratxaga, I., Sanchez, J., Navas, E., Luengo, I.: Use of the harmonic phase in speaker recognition. In: Proc. INTERSPEECH, pp. 2757–2760. Florence, Italy (2011)
14. Jaakkola, T., Haussler, D.: Exploiting generative models in discriminative classifiers. In: Proc. NIPS, pp. 487–493. Denver, CO (1999)
15. Mowlaee, P., Saeidi, R., Stylanou, Y.: INTERSPEECH 2014 special session: Phase importance in speech processing applications. In: Proc. INTERSPEECH. Singapore (2014). 5 pages
16. Mowlaee, P., Saiedi, R., Martin, R.: Phase estimation for signal reconstruction in single-channel speech separation. In: Proc. ICSLP, pp. 1–4. Hong Kong, China (2012)
17. Murthy, H., Gadde, V., et al.: The modified group delay function and its application to phoneme recognition. In: Proc. ICASSP, vol. 1, pp. I–68. Hong Kong, China (2003)
18. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: Proc. CVPR, pp. 1–8. Minneapolis, MN (2007)
19. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Proc. ECCV, pp. 143–156. Crete, Greece (2010)
20. Pittermann, J., Pittermann, A., Minker, W.: Emotion recognition and adaptation in spoken dialogue systems. International Journal of Speech Technology **13**(1), 49–60 (2010)
21. Schuller, B.: Intelligent Audio Analysis. Signals and Communication Technology. Springer (2013). 350 pages
22. Schuller, B., Batliner, A., Steidl, S., Schiel, F., Krajewski, J.: The INTERSPEECH 2011 speaker state challenge. In: Proc. INTERSPEECH, pp. 3201–3204. Florence, Italy (2011)
23. Schuller, B., Steidl, S., Batliner, A.: The INTERSPEECH 2009 emotion challenge. In: Proc. INTERSPEECH, pp. 312–315. Brighton, UK (2009)
24. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S.: The INTERSPEECH 2010 paralinguistic challenge. In: Proc. INTERSPEECH, pp. 2794–2797. Makuhari, Japan (2010)
25. Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G., Weiss, B.: The INTERSPEECH 2012 speaker trait challenge. In: Proc. INTERSPEECH. Portland, OR (2012)
26. Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., Kim, S.: The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In: Proc. INTERSPEECH, pp. 148–152. Lyon, France (2013)
27. Tahon, M., Degottex, G., Devillers, L.: Usual voice quality features and glottal features for emotional valence detection. In: Proc. ICSP, pp. 693–696. Beijing, China (2012)
28. Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., Schröder, M.: Bridging the gap between social animal and unsocial machine: A survey of social signal processing. IEEE Transactions on Affective Computing **3**(1), 69–87 (2012)
29. Wu, Z., Siong, C.E., Li, H.: Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In: Proc. INTERSPEECH. Portland, OR (2012). 4 pages
30. Xiao, X., Tian, X., Du, S., Xu, H., Chng, E.S., Li, H.: Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge. In: Proc. INTERSPEECH, pp. 2052–2056. Dresden, Germany (2015)
31. Yegnanarayana, B., Sreekanth, J., Rangarajan, A.: Waveform estimation using group delay processing. IEEE Transactions on Acoustics, Speech and Signal Processing **33**(4), 832–836 (1985)