FUSION OF MULTIPLE EMOTION PERSPECTIVES: IMPROVING AFFECT RECOGNITION THROUGH INTEGRATING CROSS-LINGUAL EMOTION INFORMATION

Chun-Min Chang and Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University, Taiwan

ABSTRACT

Developing cross-corpus, cross-domain, and cross-language emotion recognition algorithm has becoming more prevalent recently to ensure the wide applicability of robust emotion recognizer. In this work, we propose a computational framework on fusing multiple emotion perspectives by integrating cross-lingual emotion information. By assuming that each data is 'perceived' not only by a main perspective but additional derived perspectives (from a corpus of a different language), we can then combine each of the perspective-dependent features via kernel fusion technique. In specifics, we utilize two emotional corpora of different languages (Chinese and English). Our experiments demonstrate that our proposed framework achieves significant improvement over single perspective baseline across both databases.

Index Terms: speech emotion recognition, cross language, multi-task learning, affective computing

1. INTRODUCTION

Over the past decade, researchers have demonstrated the feasibility of obtaining robust emotion recognition accuracy using different measurable behavior modalities, e.g., facial expressions [1], physiology [2], speech [3], and multimodal behaviors [4, 5, 6]. Many engineering applications also benefit from the use of emotion recognition technology, e.g., in the design of natural human-computer interaction systems[7, 8, 9, 10], health care[11], marketing[12], and robotic design [13]. Speech is considered to be one of the most easily assessable data and is also the most natural form of human communication. Recently, researchers have started to investigate advanced algorithms to further improve the robustness of speech-based emotion recognition to ensure its wide applicability; this includes dealing especially with cross-corpus, cross-domain, and even cross-language scenarios.

Cross-corpus speech emotion recognition have been developed using unsupervised methods [14, 15, 16] and also based on speaker-dependent feature normalization methods [17]. Further, Zhang *et al.* have demonstrated the feasibility of performing cross-domain emotion recognition by leveraging joint characteristics in human's speech and singing in a multi-task learning framework [18]. In terms of cross-language emotion recognition, Feraru *et al.* have analyzed eight different languages, e.g., German, Danish, English, Spanish, Romanian, Turkish, Mandarin, and Burmese, in details about the transferability of conventional acoustic features when learning to recognize emotion across languages [19]. Elbarougy *et al.* have devised a three-layer model with an aim of obtaining a robust model trained on one language when applying to another language (German and Japanese) [20].

A recent meta analysis work done by Scherer *et al.* have indicated that there indeed is a substantial amount of evidences suggesting an universality of vocal emotion perception across different cultures; it still remains to be investigated though for vocal emotion production due to limited amount of relevant data [21]. Contrast to past works in cross-language emotion recognition where researchers try to study the transferability of one language to another language, our aim is to *integrate* other language useful emotion information to enhance recognition rate of the current data. We propose a novel recognition framework that works by integrating multiple *emotion perspectives*. The framework assumes that each speech sample can be 'perceived' not only by a main perspective (original label) and but additional derived perspectives (other labels). With multiple labels, this problem can then be cast as a multi-task learning where the final recognition system of the original label is trained by integrating perspective-dependent features via a multi-task kernel fusion technique.

Specifically in this work, we utilize two emotion corpora of different languages. One of them is an English corpus, i.e., the USC CreativeIT database (CIT) [22], and another one is a newly-collected Chinese corpus, the NTUA Emotion database (NTUA). Our proposed framework obtains accuracies of 0.577 and 0.507 of activation and valence dimension respectively for the CIT database and 0.682 and 0.604 of activation and valence dimension respectively for the NTUA database - both shows significant improvement compared to single perspective baselines. Further analysis demonstrates that prosodic features remain to be important across all emotion perspectives, but each perspective does bring in different aspects that benefit the overeall recognition system.

2. RESEARCH METHODOLOGY

2.1. Emotion Corpora

We utilize two similarly-collected emotion corpora, the USC CreativeIT database and the NTUA Emotion database. Table

Thanks to MOST Taiwan for funding (103-2218-E-007 -012 -MY3)

Table 1. Summary information of the USC CreativeIT and the NTUA Emotion Databases

Corpus	Language	Actors	Raters	Labels	Data
CIT	English	16	$\geq$3	VAD	90
NTUA	Mandarin	44	42	VA	204

1 summarizes key information of the two corpora.

2.1.1. The USC CreativeIT Database

The USC CreativeIT database is a publicly-available emotion corpus includes dyadic improvisations based on an established theatrical acting technique, termed the Active Analysis, in order to help elicit natural affective interactions [22]. The database consists of 16 actors (8 men and 8 women) grouped in pairs to engage in approximately 3-minute long face-to-face interactions with a 50 total interaction sessions. Audio recordings of each actor from lapel microphones is available for each actor. Each session include annotation of session-level emotion attributes of activation, valence, and dominance on a scale between $[1, 5]$ by at least 3 raters for each actor. In our experiments, we consider the average of the ratings as our emotion labels and focus only on activation and valence; there is a total of 90 samples (due to missing audio recordings of 10 samples), and each of the 90 audio recording has been previously segmented manually into utterances.

2.1.2. The NTUA Emotion Database

The NTUA Emotion database is a newly-collected Chinese emotion corpus using similar setup as the USC CreativeIT database. The corpus is collected by collaborating with Department of Drama at National Taiwan University of Arts. The database consists of 22 pair of actors grouped in dyad to engage in approximately 3-minute face-to-face interaction to act out pre-specified emotion scenarios. The scenarios are designed by the professional directors to ensure a natural and spontaneous elicitation of an overall scenario-targeted affect, i.e., happy, sad, neural, angry, surprise, and frustration. Each session includes audio recordings of each actor from lapel microphone. Each actor within each session is annotated with emotion attributes (session-level) of activation and valence on a scale between $[1, 5]$ by 42 raters. In our experiments, we consider the average of the ratings as our emotion labels, and there is a total of 204 samples - each of the audio recording has also been previously segmented into utterances.

2.2. Acoustic Feature Extraction and Encoding

In this work, we extract a high-dimensional vector to represent the acoustic profile of each actor at the session level. We first extract 45 low-level acoustic descriptors, i.e., 13 MFCCs, 1 pitch, 1 intensity, and their delta and delta-delts, every 16.6ms. We then perform Fisher-vector encoding, i.e., GMM-based encoding technique developed mainly for computer vision applications [23]. Past works have also demonstrated its effectiveness in automatic speech analyses of paralinguistic information[24, 25]. We first define a scoring function:

Fig. 1. *(Left) the CIT database (Right) the NTUA datbase*

$$G_\lambda^X = \nabla_\lambda log u_\lambda(X)$$

where $u_\lambda(X)$ denotes the likelihood of X given the probability distribution function (PDF).We use Gaussian Mixture Model (GMM) as our PDF. λ represents the parameters of GMM,$\lambda = w_k, u_k, \sum_k, k = 1, ..., K.G_\lambda^X$ is the direction where λ has to move to provide a better fit between u_λ and X. Fisher vector encoding is a special case of Fisher Kernel and its first and second order statistics below,

$$g_{u_k}^X = \frac{1}{T\sqrt{w_k}} \sum_{t=1}^{T} \gamma_t(k) \left(\frac{x_t - u_k}{\sigma_k} \right)$$

$$g_{\sigma_k}^X = \frac{1}{T\sqrt{2w_k}} \sum_{t=1}^{T} \gamma_t(k) \left(\frac{(x_t - u_k)^2}{\sigma_k^2} - 1 \right)$$

$\gamma_t(k)$ is defined as

$$\gamma_t(k) = \frac{w_k u_k(x_t)}{\sum_{j=1}^{K} w_j u_j(x_t)}$$

where $w_k, u_k, \sum_k, k = 1, ..., K$ correspond to mixture weight, mean, and covariance matrix for each mixture of Gaussian. In specifics, we use 128-GMM and retrieve the mean and variance as our final encoded features for each data sample (dimension = $45 \times 2 \times 128 = 11520$) in this work.

2.3. Fusion of Multiple Emotion Perspectives

The fusion of multiple emotion perspectives is designed such that each of the speech sample is assumed to have been *"perceived"* by a main perspective (original emotion label) and two additional perspectives (generated from the other corpus). The final system recognizes the emotion attribute of the main perspective through integrating all the perspective-dependent features via a kernel fusion technique during training. In specifics, for every data, A_i, we treat each perspective as a *label* for that sample. There are a total of three different labels (1 main and 2 derived):

1. **Main-Perspective**: the original affect dimension labels for each speech sample
2. **Derived-CosDist**: for every sample, A_i in database A, look for the most similar k samples in database B using cosine distance computed on acoustic features (section 2.2), then the derived perspective for A_i is obtained by averaging the human-rated labels on those k samples
3. **Derived-XPred**: first, train a SVM regressor, B_{recg}, using all data in B. For every sample, A_i, use B_{recg} to predict its emotion label as the derived perspective.

With respect to each of the above perspectives, we then obtain a different set of features after performing ANOVA F-test feature selection. Each perspective-dependent feature set can be imagined as the *relevant* acoustic features with respect to a

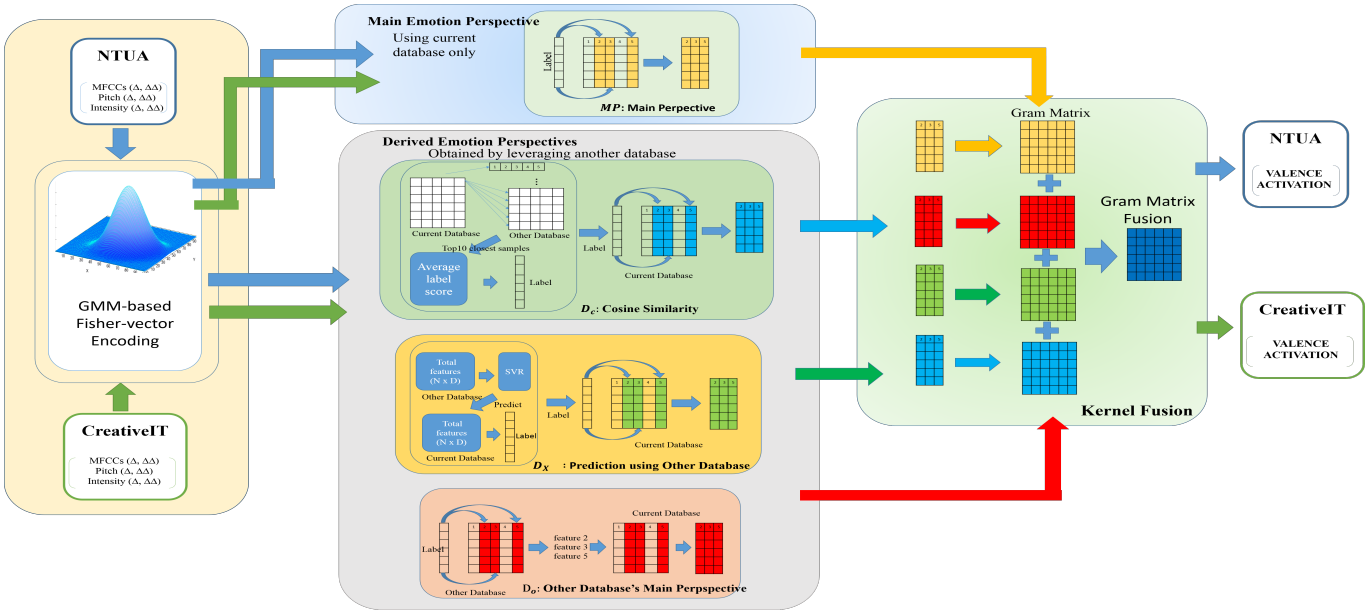

Fig. 2. *Illustration on the system architecture of our proposed multiple emotion perspective fusion framework*

particular viewing angle, e.g., *Derived-CosDist* of database A can be imagined as though raters in database B had annotated the emotion for samples in A, and so on.

The final framework is based on training linear support vector regression on kernel fused with gram matrices generated from each *perspective-dependent* feature set - an improved method that has outperformed common multi-task feature learning method [26, 27]. Given features sets $F_1, \ldots, F_i, \ldots, F_N$, we compute kernel matrix of each set:

$$K_i = k(F_i, F_i^{'}), i = 0, 1, ...N$$

then, we integrate all kernels using a function $f(\cdot)$

$$K = f(K_0, K_1, K_2, ..., K_N)$$

In this work, linear kernel is chosen for $k(\cdot)$, and simple summation ($f(\cdot)$) is used to combine the kernel matrices.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Experimental Setup

We present three different baseline emotion recognition systems ($\mathrm{B}_1, \mathrm{B}_2, \mathrm{M}_P$) for both CIT and NTUA databases. B_1 indicates that the GMM trained to derive Fisher-vector encoding feature (section 2.2) is done by using individual database only, B_2 indicates that the GMM is trained by joining both databases, and M_P is B_2 with feature selection (essentially the features derived from the main perspective method mentioned in section 2.3). For both CIT and NTUA, we conduct regression experiments using leave-dyad-out cross validation on emotion dimensions of activation and valence. The evaluation metric is spearman correlation. All the feature selection is carried out on the training set. On the multiple emotion perspective fusion, we have four different sets of features:

- $\mathbf{M}_P$: Main-Perspective selected features
- $\mathbf{D}_C$: Derived-CosDist selected features
- $\mathbf{D}_X$: Derived-XPred selected features
- $\mathbf{D}_O$: Main-Perspective features of the *other* database

The first three sets are from the three perspectives mentioned in section 2.3. Here, we add addition set D_O. It simply means that for all samples in database, A, we select features based on the feature selection conducted on the database, B; this essentially means using the main perspective selected feature of B for A. We present multi-perspective results by iteratively adding various D_* to M_P via kernel fusion. Except for B_1, the GMMs for Fisher-vector encoding is all done by training the GMMs on joining both CIT and NTUA databases.

3.2. Experimental Results and Analyses

Table 2 lists a summary of our emotion recognition results for both of the databases. The number in the parenthesis indicates the percentages of feature selected for that perspective, i.e., empirically determined to obtain the best result for each individual perspective. There are several interesting observations. The first one is that when comparing B_1 to B_2, it is evident that by simply training the GMM jointly, it already provides an improvement in the robustness of the acoustic feature representation and, hence, increases the emotion recognition accuracies across two databases on both emotion dimensions.

Secondly, the multiple emotion perspective fusion framework indeed provides additional improvement in the recognition accuracies. In specifics, for the CIT database, we achieve the best results of 0.565 and 0.507 correlation for the activation and valence respectively, which is 10.57% and 6.96% relative improvement over the single main perspective baseline (M_p); for the NTUA database, we achieve the best results of 0.682 and 0.564 correlation for the activation and valence respectively, which is 3.49% and 2.37% relative improvement

Table 2. *Summary on the emotion recognition accuracies for the two databases on dimension of activation and valence*

The USC CreativeIT Database (CIT)

Act : $M_P(80)$ $D_C(80)$ $D_X(30)$ $D_O(30)$, Val : $M_P(50)$ $D_C(70)$ $D_X(90)$ $D_O(70)$

	Baseline			$\mathbf{M}_P$			$\mathbf{M}_P$			$\mathbf{M}_P$
	$\mathbf{B}_1$	$\mathbf{B}_2$	$\mathbf{M}_P$	$+\mathbf{D}_C$	$+\mathbf{D}_X$	$+\mathbf{D}_O$	$+\mathbf{D}_C + \mathbf{D}_X$	$+\mathbf{D}_C + \mathbf{D}_O$	$+\mathbf{D}_X + \mathbf{D}_O$	$+\mathbf{D}_C + \mathbf{D}_X + \mathbf{D}_O$
Act.	0.483	0.510	0.511	0.525	0.553	0.541	0.546	0.538	**0.565**	0.553
Val.	0.341	0.466	0.474	**0.507**	0.481	0.486	0.490	0.492	0.469	0.486

The NTUA Emotion Database (NTUA)

Act : $M_P(30)$ $D_C(50)$ $D_X(80)$ $D_O(90)$, Val : $M_P(50)$ $D_C(90)$ $D_X(90)$ $D_O(90)$

	Baseline			$\mathbf{M}_P$			$\mathbf{M}_P$			$\mathbf{M}_P$
	$\mathbf{B}_1$	$\mathbf{B}_2$	$\mathbf{M}_P$	$+\mathbf{D}_C$	$+\mathbf{D}_X$	$+\mathbf{D}_O$	$+\mathbf{D}_C + \mathbf{D}_X$	$+\mathbf{D}_C + \mathbf{D}_O$	$+\mathbf{D}_X + \mathbf{D}_O$	$+\mathbf{D}_C + \mathbf{D}_X + \mathbf{D}_O$
Act.	0.633	0.649	0.659	0.679	0.681	0.676	**0.682**	0.679	0.673	0.676
Val.	0.568	0.596	0.590	0.602	0.596	**0.604**	0.596	**0.604**	0.599	0.598

over the single main perspective baseline (M_p). We observe that, in general, fusion of different perspectives always outperforms single perspective framework. This also corroborates the analyses finding that Scherer states about the possible universality of vocal emotion perception across different cultures [21]. Furthermore, we compute the inter-perspective spearman correlations (Table 3), and we observe that the correlation is low. It seems to indicate that while different perspectives may not agree with each other on the emotion labels, the relevant acoustic features that bear perceptual information with respect to each perspective, however, when integrated are actually complementary to the original features in terms of modeling the the main perspective emotion labels.

Lastly, while the features are encoded in terms of means and variances of Fisher-vector, we can still retract the type of original low level descriptors that are being selected out of the Fisher-vector encoding. We examine the top 10 out of 45 low level descriptors that are associated with the selected Fisher-vector features for each perspective. Prosody (i.e., intensity and pitch) descriptors, which account for only 13% of the total encoded feature dimensions, always shows up across all perspectives - an intuitive pleasing result in line with the past research in showing that prosody is a robust measure of emotion [28, 29]. Furthermore, Table 3 shows the actual list of prosody-related descriptors selected for the "main-perspective" and the "derived-perspectives". In general, we see that the exact descriptors are not the same across perspectives. This result further implicates that our multiple emotion perspectives framework is indeed capable of extracting and adding useful emotion-related features that contributes to the original emotion perception but are *hidden* and/or difficult to identify in the conventional single perspective approach.

Table 3. *Perspectives and Feature Analyses*

The CIT Database				**The NTUA Database**			
Act.		Val.		Act.		Val.	
Inter-Perspective Correlation							
(M_P vs. D_C) / (M_P vs. D_X)							
0.22 / 0.19		0.08 / 0.17		0.30 / 0.28		0.02 / 0.09	
Feature Selection Analysis							
M_P	D_*	M_P	D_*	M_P	D_*	M_P	D_*
$Int_{\Delta\Delta}$	F0	F0	F0	$F0_{\Delta\Delta}$	F0	$F0_{\Delta\Delta}$	F0
Int_{Δ}	$F0_{\Delta}$		$F0_{\Delta}$		$F0_{\Delta}$	Int	Int
	Int		$F0_{\Delta\Delta}$		$F0_{\Delta\Delta}$	Int_{Δ}	Int_{Δ}
			Int		Int	$Int_{\Delta\Delta}$	$Int_{\Delta\Delta}$
			Int_{Δ}		Int_{Δ}		
			$Int_{\Delta\Delta}$		$Int_{\Delta\Delta}$		

4. CONCLUSIONS

In this work, we propose a novel framework of fusing multiple emotion perspectives, i.e., derived from a different language, to improve the emotion recognition system compared to conventional single language emotion recognizer. Through integrating cross-lingual emotion information, we achieve a significant improvement in both activation and valence dimensions across two databases. Furthermore, our analyses indicate that our framework has the potential of extracting relevant acoustic features relating to emotion perception that may be hidden if given only a single language database; these promising results also seems to corroborate with the theorized cross-cultural universality on vocal emotion perception.

There are several future directions. One of the immediate future directions is to extend the framework to incorporate multiple language families and to extract a more variety of acoustic parameters, e.g., voice qualities. Second, the key components of our framework involves semi-supervised learning together with multi-task learning, we will investigate algorithm to jointly optimize these two current independent components to further improve the robustness. Third, the interactions included in these two databases are long-duration in nature, our past work has demonstrated the effectiveness of identifying emotional thin-slices in performing emotion recognition[25]. We plan on incorporating such a concept into this proposed framework. Lastly, on the long term, we hope to advance the knowledge on understanding underlying mechanism of cross-culture vocal emotion perception and substantiate the theory of universality of vocal emotion perception.

5. REFERENCES

[1] Michel F Valstar, Marc Mehu, Bihan Jiang, Maja Pantic, and Klaus Scherer, "Meta-analysis of the first facial expression recognition challenge," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 966–979, 2012.

[2] Jonghwa Kim and Elisabeth André, "Emotion recognition based on physiological changes in music listening," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[3] Chul Min Lee and Shrikanth S Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE transactions on speech and audio processing*, vol. 13, no. 2, pp. 293–303, 2005.

[4] Yelin Kim, Honglak Lee, and Emily Mower Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 3687–3691.

[5] Zhihong Zeng, Jilin Tu, Brian M Pianfetti, and Thomas S Huang, "Audio–visual affective expression recognition through multistream fused hmm," *IEEE Transactions on Multimedia*, vol. 10, no. 4, pp. 570–577, 2008.

[6] Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 2462–2465.

[7] Alan Dix, *Human-computer interaction*, Springer, 2009.

[8] Paul Bach-y Rita and Stephen W Kercel, "Sensory substitution and the human–machine interface," *Trends in cognitive sciences*, vol. 7, no. 12, pp. 541–546, 2003.

[9] Soon Heung Chang, Seong Soo Choi, Jin Kyun Park, Gyunyoung Heo, and Han Gon Kim, "Development of an advanced human–machine interface for next generation nuclear power plants," *Reliability Engineering & System Safety*, vol. 64, no. 1, pp. 109–126, 1999.

[10] Deborah L Pinard, Eliana MO Peres, and Ronald A Evans, "Human machine interface for telephone feature invocation," July 2 1996, US Patent 5,533,110.

[11] Christina Lisetti, Fatma Nasoz, Cynthia LeRouge, Onur Ozyer, and Kaye Alvarez, "Developing multimodal intelligent affective interfaces for tele-home health care," *International Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 245–255, 2003.

[12] Fuji Ren and Changqin Quan, "Linguistic-based emotion analysis and recognition for measuring consumer satisfaction: an application of affective computing," *Information Technology and Management*, vol. 13, no. 4, pp. 321–332, 2012.

[13] Neville Hogan, Hermano I Krebs, Andre Sharon, and Jain Charnnarong, "Interactive robotic therapist," Nov. 14 1995, US Patent 5,466,213.

[14] Zixing Zhang, Felix Weninger, Martin Wöllmer, and Björn Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 523–528.

[15] Daniel Bone, Chi-Chun Lee, and Shrikanth Narayanan, "Robust unsupervised arousal rating: A rule-based framework withknowledge-inspired vocal features," *IEEE transactions on affective computing*, vol. 5, no. 2, pp. 201–213, 2014.

[16] Jun Deng, Zixing Zhang, Florian Eyben, and Björn Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.

[17] Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wollmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll, "Cross-corpus acoustic emotion recognition: variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[18] Biqiao Zhang, Emily Mower Provost, and Georg Essi, "Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5805–5809.

[19] Silvia Monica Feraru, Dagmar Schuller, et al., "Cross-language acoustic emotion recognition: An overview and some tendencies," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 125–131.

[20] Reda Elbarougy and Masato Akagi, "Cross-lingual speech emotion recognition system based on a three-layer model for human perception," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*. IEEE, 2013, pp. 1–10.

[21] Klaus R Scherer, Elizabeth Clark-Polner, and Marcello Mortillaro, "In the eye of the beholder? universality and cultural specificity in the expression and perception of emotion," *International Journal of Psychology*, vol. 46, no. 6, pp. 401–435, 2011.

[22] Angeliki Metallinou, Zhaojun Yang, Chi-chun Lee, Carlos Busso, Sharon Carnicke, and Shrikanth Narayanan, "The usc creativeit database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations," *Language resources and evaluation*, pp. 1–25, 2015.

[23] Ken Chatfield, Victor S Lempitsky, Andrea Vedaldi, and Andrew Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods.," in *BMVC*, 2011, vol. 2, pp. 8–19.

[24] Heysem Kaya, Alexey A Karpov, and Albert Ali Salah, "Fisher vectors with cascaded normalization for paralinguistic analysis," *Proc. of INTERSPEECH. Dresden, Germany: ISCA*, pp. 909–913, 2015.

[25] Wei-Cheng Lin and Chi-Chun Lee, "A thin-slice perception of emotion? an information theoretic-based framework to identify locally emotion-rich behavior segments for global affect recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5790–5794.

[26] Edwin V Bonilla, Felix V Agakov, and Christopher KI Williams, "Kernel multi-task learning using task-specific features.," in *AISTATS*, 2007, pp. 43–50.

[27] Andrea Caponnetto, Charles A Micchelli, Massimiliano Pontil, and Yiming Ying, "Universal multi-task kernels," *Journal of Machine Learning Research*, vol. 9, no. Jul, pp. 1615–1646, 2008.

[28] Carlos Busso, Sungbok Lee, and Shrikanth Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 582–596, 2009.

[29] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al., "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.