A Bootstrapped Multi-View Weighted Kernel Fusion Framework for Cross-Corpus Integration of Multimodal Emotion Recognition

Chun-Min Chang
Department of Electrical Engineering
National Tsing Hua University, Taiwan
cmchang@gapp.nthu.edu.tw

Bo-Hao Su
Department of Electrical Engineering
National Tsing Hua University, Taiwan
bo0918990693@gapp.nthu.edu.tw

Shih-Chen Lin
Department of Electrical Engineering
National Tsing Hua University, Taiwan
dennis60512@gapp.nthu.edu.tw

Jeng-Lin Li
Department of Electrical Engineering
National Tsing Hua University, Taiwan
cllee@gapp.nthu.edu.tw

Chi-Chun Lee
Department of Electrical Engineering
National Tsing Hua University, Taiwan
cclee@ee.nthu.edu.tw

***Abstract*—Recently the development of robust emotion recognition has been increasingly emphasized in order to handle situations of different cultures and languages. This has become critical due to the potential applicability of emotion recognizers across a wide range of application scenarios. Instead of conventional approach in deriving a single universal emotion recognition module across all languages, we have previously demonstrated a method based on integrating other database's useful information to improve the emotion recognition of the current data with fusion of *multiple emotion perspectives*. In this paper, we present an improved framework, i.e., a bootstrapped multi-view weighted kernel fusion, to further advance the recognition accuracies. We have also extended the modeling of speech-only modality to include video information. In specifics, we utilize two emotional corpora of different languages. Our proposed framework obtains improved recognition in regressing activation and valence attributes using audio and video modalities across both of the databases. We not only demonstrate that the weighted kernel fusion can provide additional modeling power but also present analyses on the complementary emotionally-relevant acoustic and visual behaviors computed from the multiple emotion perspectives.**

1. Introduction

Continuous algorithmic advancement in affective computing has positioned automated emotion sensing capability to become an essential module across applications, e.g., natural human-computer interface [1], health care [2], marketing [3], and robotic design [4]. While there has been a tremendous technical development, obtaining robust emotion recognition results especially in contexts of cross-language, culture, and corpus, remains to be a challenging yet critical issue. Previous works in cross-corpus emotion recognition concentrate on developing frameworks either to achieve robust transferability or to obtain an universal emotion recognizer across settings. Some exemplary works include the use of unsupervised (e.g., knowledge-driven) methods in obtaining robustness across multiple databases [5], [6]. Speaker dependent normalization has also been introduced as a plausible strategy in tasks of cross-copora emotion recognition [7]. Zhang et al. have indicated the feasibility of leveraging joint characteristics in cross-domain, i.e., speech and singing, emotion recognition with a multi-task learning framework [8]. Lastly, 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 in recognizing emotion across languages [9].

A recent meta analysis work done by Scherer et al. shows that there are substantial psychological evidences indicating the common universality in vocal emotion perception across cultures and languages. However, most of past computational works have seen that due to potentially a high variability in the recording conditions, labeling definitions, and other idiosyncratic factors, obtaining a reliable single universal recognizer remains to be extremely challenging. Therefore, our approach differs in that it is based on *integrating* other language's complementary emotion information to further enhance the recognition accuracy of the current language. In fact, we have previously proposed a method based on fusion of *multiple emotion perspectives*. By assuming each sample within the current emotion database can be perceived with multiple diverse perspectives (labels), e.g., main perspective (manual label originally given in the database) and derived perspectives (artificial labels derived from another database), we can then carry out a multi-view fusion framework to leverage the joint emotionally-relevant information between these different perspective. In fact, our previous paper demonstrates an improved speech emotion recognition accuracies by integrating these different emotion perspectives with a kernel-fusion framework [10].

The use of multi-view fusion technique is naturally appealing in this task since the core idea is to combine

978-1-5386-0563-9/17/$31.00 ©2017 IEEE

IEEE computer society

TABLE 1. SUMMARY INFORMATION OF THE USC CREATIVEIT AND THE NNIME EMOTION DATABASES

Corpus	Language	Actors	Raters	Labels	Data
CIT	Eng.	16	≥ 3	VAD	89
NNIME	Mand.	44	42	VA	201

decisions on recognizing an attribute from different angles, where each view presents a fractional and complementary information on the attribute-of-interest (in this case, emotion state). For example, Xu et al. proposes a large-margin multi-view information bottleneck algorithm with multiple senders in a communication system, where each of which represents one view of the data [11]. Dhillon et al. utilizes multi-view learning in terms of low rank learning to estimate low dimensional context-specific word representations from different sources of unlabeled data [12]. Sun also provides a summary on current progresses in multi-view learning [13].

In our previous work of speech-based cross-corpus integrative emotion recognition [10], we first generate multiple emotion perspectives for each of the database, and the multi-view fusion is carried out on learning a fused kernel matrix from perspective-dependent features. In this paper, we extend upon that previous framework with two novel contributions: 1) introducing the use of bootstrapped weighting on selected perspective-dependent features in the process of learning the fused kernel matrix, and 2) extending the speech-only modality to multimodal behaviors with inclusion of video information. Incorporating bootstrapped weighting for fused-kernel learning encodes the importance of these emotionally-relevant behavior features within each view and subsequently improves the robustness of our framework. Further, while multimodal behaviors have been shown to be important in emotion expressions (e.g., [14], [15]), to the best of our knowledge, there has not been works done in modeling body language from video in settings of cross-corpora emotion recognition.

In this work, we utilize two emotion corpora: one of them is an English database, the USC CreativeIT database (CIT) [16], and a Chinese database, the NTHU-NTUA Chinese Interactive Multimodal Emotion Corpus (NNIME) [17]. Our proposed framework obtains a correlation of 62.3%, 31.6% and 55.4%, 50.5% in regressing activation and valence dimensions using audio and video modality respectively for the CIT database. For the NNIME database, we obtain a correlation 71.3%, 50.0% and 59.9%, 49.0% for regressing activation and valence dimensions using audio and video modality respectively. In both corpora, we achieve an improvement over the previously-proposed framework [10], and we further provide analyses on the complementary nature of different emotion perspectives. The rest of paper is organized as follows: section 2 describes about the two databases; section 3 details our proposed framework; section 5 details our experimental setup, results, and analyses; finally, section 5 concludes with future works.

2. Emotion Databases

We utilize two similarly-collected multimodal emotion corpora, the USC CreativeIT database and the NNIME database in this work. Table 1 summarizes key information of the two databases, and Figure 1 shows a snapshot of the actual recording session of each database.

Figure 1. *A snapshot of the two databases used: (left) the CIT database (right) the NNIME datbase*

2.1. The USC CreativeIT Database (CIT)

The USC CreativeIT database (CIT) is a publicly-available emotion corpus consists of dyadic interactions. It features the use of a novel theatrical technique, i.e., the active analysis, to elicit natural emotions and expressive speech and body language during interactions [16]. There are 16 actors (8 men and 8 women) in the database that are divided into 8 pairs to perform 3 to 5 minutes long improvised actings, This results in a total of 50 sessions in the database. For every session, multimodal behavior data of both subjects are collected, which include audio recordings of lapel microphones and video recordings of a camera standing at the side of the recording space.

The perceived emotions of each actor are annotated at the session-level using the dimensional attributes of activation and valence on a scale between $[1, 5]$ by 3 rates. In this work, we conduct our recognition experiments using 89 samples due to 9 missing audio samples and 2 problematic video samples. The ground true labels of our work are the average values of the three raters. Note that all audio files have been previously segmented manually into utterances, and we design a tracking algorithm to locate the bounding box of each individual actor's in the recorded video stream.

2.2. The NNIME Emotion Database (NNIME)

The NNIME emotion database is a novel Chinese dyadic interaction corpus which results from the collaborative work between engineers and drama experts [17]. It adopts a similar emotional behavior elicitation setup as the CIT database. The NNIME database consists of 22 distinct pairs of actors grouped in dyad to participate in approximately 3-minute long affective interactions. The design of interactions is to emulate real-life scenarios with an overall target emotional plot, and a professional director is involved to ensure the naturalness and qualities of the interactions. For each session, a video camera facing the stage is used to record the movement of the dyad, and audio information is recorded using lapel microphone placed on individual actors.

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 use the average of the ratings as our ground truth emotion labels. While there are a total of 204 annotated samples, 3 rout of 204 are recorded badly. 201 samples are used as our dataset. The audio files of the NNIME are pre-segmented manually, and the actors' video are pre-processed by our proposed tracking algorithm to identify the bounding box of each individual in every video frame.

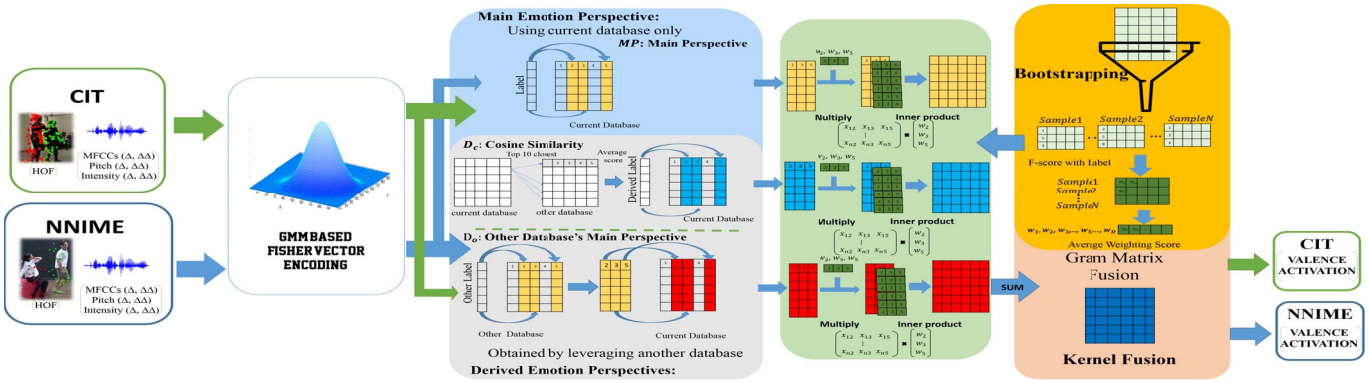

Figure 2. *Illustration on the system architecture of our proposed multiple emotion perspective fusion framework: extraction of acoustic and visual low level behavior descriptors, learning of high-dimensional behavior representation using GMM Fisher-vector, derivation of multiple emotion perspectives, constructing fused kernel matrix with bootstrapping feature weights to be used in the final support vector regression of activation and valence dimension in the USC CreativeIT and the NNIME emotion databases*

3. Multimodal Emotion Recognition

Figure 2 displays our overall recognition architecture based on novel fusion of multiple emotion perspectives. In this section, we will describe each component of the architecture: extracting acoustic and visual low level descriptors, learning high-dimensional behavior representation for each individual at the session level, deriving multiple emotion perspectives for each sample, constructing fused kernel matrix to be used in the final support vector regression.

3.1. Behavior Low-level Descriptors (LLDs)

3.1.1. Acoustic LLDs. Since both corpora provide segmented speaking terms for each actor, we extract 45 low-level descriptors characterizing frame-level acoustic features on the entire speaking duration of each subject. The LLDs include 13 mel-frequency cepstral coefficients (MFCCs), 1 fundamental frequency, 1 intensity, and their delta and delta-delta computed at 60 frames per second.

3.1.2. Visual LLDs. For both corpora, we first implement a method based on computing histogram of oriented gradient (HOG) with temporal displacement constraint in order to detect and track an individual in the video stream. In case of occlusion, if a person's bounding box disappears and reappears, that bounding box is likely to belong to the one being occluded with additional constraint placed on the absolute displacement of the reappearing bounding box. This tracking algorithm achieves about 90% accuracies in reliably identifying each subject's bounding box at every frame in the video stream across both of the databases.

We extract 5 motion frame-level low-level descriptors within the located bounding box. The 5 descriptors are trajectories, histogram of oriented gradient (HOG), histogram of oriented optical flow (HOF), and motion boundary histogram (MBH). In this work, through our empirical experiments, we decide to use HOF as the final visual low-level descriptors. An important advantage of computing optical flow is that it imposes a temporal smoothing constraint when tracking moving pixels in the video sequences. HOF is a 108 dimensional feature vector. For a frame t, it describes motion at 3 different time points, $t-1, t, t+1$, each with 36 dimensions; 36 dimensions are computed from splitting the bounding box into 4 equal-sized spatial region in which 9 bins of histogram are calculated to summarize the movement in 9 different directions for each spatial region.

3.2. Session-level Behavior Representation

To encode varying-length frame-level LLDs to fixed-dimension representations, we further use Gaussian Mixture Model based Fisher Vector (GMM-FV) encoding approach to learn the session-level behavior representation of each modality. A brief description of GMM-FV is given below, for a data sequence X, we can define a scoring function:

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

where $u_\lambda(X)$ denotes the likelihood of X given the probability distribution function (PDF). We use GMM as our PDF. λ represents the parameters of GMM, $\lambda = w_k, u_k, \sum_k, k = 1, ..., K$. G_λ^X is the direction where λ has to move to provide a better fit between u_λ and X. Fisher vector encoding is derived by computing the following first and second order statistics:

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

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

$\gamma_t(k)$ 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 GMM with $M = 128$ for both audio and video modalities to derive the session-level features of each data sample. This is the feature input used in the following weighted kernel fusion.

3.3. Fusion of Multiple Emotion Perspectives

The center idea of our computational framework is to leverage the joint emotionally-relevant behavior information derived from having access to the two different emotion

corpora (denoted as A & B). Specifically, if we assume each sample of a database can be perceived by different emotion perspectives, then we can fuse the diverse *perspective-dependent* features to better model the complex behavior manifestations of the original target emotion attribute. In this work, for every sample i in each database, there are two emotion perspectives. For example, given a particular sample A_i, the two perspectives are: 1) *Main-Perspective*: the original human-rated affect labels, and 2) *Derived-Perspective*: using cosine distance to find the closest n behavior samples in B and averaging the human-rated affect labels of those behavior samples in B to be the additional label placed on A_i ($n = 10$ in this case).

With these perspectives, we can obtain three distinct sets of *perspective-dependent* features for each sample i using the bootstrapped feature selection method (section 3.3.1):

- $\mathbf{M}_P$: Features outputted from selecting with respect to *Main-Perspective* of the current database
- $\mathbf{D}_C$: Features outputted from selecting with respect to *Derived-Perspective*
- $\mathbf{D}_O$: Features outputted from selecting with respect to *Main-Perspective* of the other database

We then integrate these diverse cross-corpus information by computing a weighted fused kernel matrix to be used as the kernel for support vector regression. We will describe this framework in learning the weighted fused-kernel below.

3.3.1. Bootstrapping Feature Importance Weighting. The three sets of dependent features are selected based on ANOVA F-test. We additionally use a meta bootstrapping method in the selection procedure to improve the robustness. The procedure is listed below:

First, we randomly sample M data out of the entire database, then given the training vectors $x_k, k = 1, 2, ..., M$ (each vector is D dimension) with perspective-dependent label $t_k, k = 1, 2, ..., M$, we compute the following:

$$Corr(i) = \sum_{k=1}^{M} \frac{(x_{k,i} - \bar{x_i}) \cdot (t_k - \bar{t})}{\sqrt{\sum_{k=1}^{M}(x_{k,i} - \bar{x_i})^2} \times \sqrt{\sum_{k=1}^{M}(t_k - \bar{t})^2}}$$

where $i = 1, 2, ..., D$. We then compute Fscore, which indicates the importance of each feature i, defined as:

$$Fscore(i) = \frac{Corr(i)^2}{1 - Corr(i)^2}$$

By repeating the above procedure P times (the bootstrapped iteration), we obtain the importance weighting of each feature i as the average of Fscore defined below:

$$w(i) = \frac{\sum_{p=1}^{P} Fscore_{p,i}}{P}$$

Having computed $w(i)$, we not only obtain the criteria to select the *perspective-dependent* features but also their associated importance weighting at the same time.

3.3.2. Weighted Kernel Fusion. We use kernel fusion method to integrate multiple perspectives of emotion information. The method is based on constructing a fused gram matrix from different feature sets to be used in the support vector machine, i.e., given features sets $F_1, \ldots, F_i, \ldots, F_N$, one first computes kernel matrix of each set:

$$K_j = k(F_j, F_j^{'}), j = 0, 1, ...N$$

then integrates all kernels using a function $f(\cdot)$

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

In this work, we incoporate the use of weights computed in section 3.3.1:

$$F_j = (w \cdot x)_j, j = 0, 1, ...N$$

$$K = \sum_{j=1}^{N} k((w \cdot x)_j, (w \cdot x)_j^{'}), j = 0, 1, ...N$$

Linear kernel is chosen for $k(\cdot)$, and K is final fused-kernel matrix integrating multiple emotion perspectives information across corpus to be used in the training of the support vector regression.

4. Experimental Setup and Results

4.1. Experimental Setup

We first compare our algorithms to three different baseline systems (B_1,B_2,M_P), i.e, ones with single perspective only, for both the NNIME and the CIT in acoustic and video modalities. B_1 indicates that LLD features have been encoded with GMM-FV, where the GMM is trained on each individual database, B_2 indicates that the GMM is trained with both databases. M_P, i.e., the *Main-Perspective* is B_2 with conventional feature selection (ANOVA F-test without the bootstrapping). For both the CIT and the NNIME, we use leave-one-dyad-out cross validation for activation and valence regression experiment on acoustic and visual modality separately. The evaluation metric is spearman.

We further compare our bootstrapped weighted kernel fusion with a the fusion framework without the weighting, i.e., the previous published work [10]. Hence, there will essentially be two kinds of kernel fusion:

- $\mathbf{M}_P + (\mathbf{D}_C$ or $\mathbf{D}_O)$: Integration of *Main-Perspective* with one of the *Derived-Perspective*
- $\mathbf{M}_{Pw} + (\mathbf{D}_{Cw}$ or $\mathbf{D}_{Ow})$: Weighted integration of *Main-Perspective* with a *Derived-Perspective*

4.2. Experimental Results and Analyses

Table 2 lists a summary of our emotion recognition results experiments for both speech and video modality on the CIT and the NNIME emotion database. For the CIT database, the best speech-based results are obtained by using weighted fusion of multiple emotion perspectives ($M_{pw} + D_{*w}$), and we achieve correlations of 0.623 and 0.554 for activation and valence respectively, which is an 6.9% and 10.3% relative improvement over the best single-perspective baseline, M_p. In the video modality, our proposed fusion method also obtains the best results, i.e., 0.316 and 0.505 for activation and valence (13.2% and 12.6% relative improvement over it respective M_p).

For the NNIME database, we observe similar boosts in accuracies by fusing emotion perspectives. In specifics, in the speech modality, the best results obtained are 0.713

TABLE 2. *Summary on the multiple fusion of emotion perspectives recognition accuracies for the two databases on dimension of activation and valence. The metric of choice is Spearman correlation. Baselines are single perspective results.* $\mathbf{M}_P$, $\mathbf{M}_{Pw}$, $\mathbf{D}_*$, *and* $\mathbf{D}_w$ *indicate Main-Perspective and Derived-Perspective without or with bootstrapped weighting*

The USC CreativeIT Database (CIT)														
	speech (s)							**video (v)**						**s+v**
	Baseline			$\mathbf{M}_P+$		$\mathbf{M}_{Pw}+$		**B**aseline			$\mathbf{M}_P+$		$\mathbf{M}_{Pw}+$	**A**vg
	$\mathbf{B}_1$	$\mathbf{B}_2$	$\mathbf{M}_P$	$\mathbf{D}_C$	$\mathbf{D}_O$	$\mathbf{D}_{Cw}$	$\mathbf{D}_{Ow}$	$\mathbf{B}_1$	$\mathbf{B}_2$	$\mathbf{M}_P$	$\mathbf{D}_C$	$\mathbf{D}_O$	$\mathbf{D}_{Cw}$	$\mathbf{D}_{Ow}$	
Act.	0.483	0.554	0.554	0.565	0.589	0.592	**0.623**	0.037	0.184	0.184	0.189	0.183	0.276	**0.316**	0.514
Val.	0.341	0.451	0.451	0.468	0.497	**0.554**	0.526	0.266	0.327	0.379	0.358	0.378	0.462	**0.505**	0.601
The NNIME Emotion Database(NNIME)														
	speech (s)							**video (v)**						**s+v**
	Baseline			$\mathbf{M}_P+$		$\mathbf{M}_{Pw}+$		**B**aseline			$\mathbf{M}_P+$		$\mathbf{M}_{Pw}+$	**A**vg
	$\mathbf{B}_1$	$\mathbf{B}_2$	$\mathbf{M}_P$	$\mathbf{D}_C$	$\mathbf{D}_O$	$\mathbf{D}_{Cw}$	$\mathbf{D}_{Ow}$	$\mathbf{B}_1$	$\mathbf{B}_2$	$\mathbf{M}_P$	$\mathbf{D}_C$	$\mathbf{D}_O$	$\mathbf{D}_{Cw}$	$\mathbf{D}_{Ow}$	
Act.	0.666	0.661	0.667	0.682	0.691	0.704	**0.713**	0.468	0.469	0.469	0.436	0.447	**0.500**	0.498	0.695
Val.	0.567	0.564	0.564	0.562	0.564	0.596	**0.599**	0.297	0.414	0.415	0.441	0.436	0.470	**0.490**	0.598

and 0.599 for activation and valence, corresponding to 4.6% and 3.5% relative improvements over the M_p. In the visual modality, the best accuracies are 0.50 and 0.49 that improves 3.1% and 7.5% relatively over the M_p. In fact, we observe that even just by using the GMM trained with both databases in the process of deriving GMM-FV, we see an improvement in the recognition rates compared to using GMM trained solely on an individual database (B_1 versus M_p). This is evident across emotion attributes, behavior modalities, and databases (only exception is the speech modality in the NNIME). These results demonstrate the effectiveness of our core idea, i.e., by integrating joint emotionally-relevant information across databases, we can improve the emotion recognition of the current data.

Furthermore, by comparing accuracies obtained when using our proposed bootstrapped weighted kernel fusion in this work versus past, i.e., no weighting, kernel fusion ($\mathrm{M}_p + \mathrm{D}_*$ versus $\mathrm{M}_p w + \mathrm{D}_* w$), we observe that the proposed weighting indeed provide substantial improvements. In specifics, for the CIT database, the accuracy increases from 0.589 to 0.623 in activation and 0.497 to 0.554 in valence for speech modality; in the video modality, the accuracies improve from 0.189 to 0.316 in activation and 0.378 to 0.505 in valence. Similarly for the NNIME database, the accuracy obtained using speech features increases from 0.691 to 0.713 in activation and 0.564 to 0.599 in valence and further raises up from 0.447 to 0.500 in activation and 0.441 to 0.490 in valence when using video features. A simple mulitmodal fusion by averaging the regressed values of audio and video can further improve the recognition accuracy in the valence dimension of the CIT to over 0.6. Our newly-proposed bootstrapped weighting (section 3.3.1 and 3.3.2) takes the importance of each *perspective-dependent* features into account and improve the robustness of our overall fusion architecture of multiple emotion perspectives.

4.2.1. Additional Analysis. Since by weighted integration of *Main-Perspective* and *Derived-Perspective*, we obtain significant improvements. We further perform analyses on the types of features that each perspective bring to the emotion recognition system. While the features are encoded in the high-dimensional GMM-FV representations, we can back-trace to identify which original LLD that each dimension in the GMM-FV space refers to. Figure 3 shows a plot where x-axis indicates the list of all of the original LLDs and y-axis shows the distribution on the original LLDs out of the top 25% weighted GMM-FV dimensions selected (section 3.3.1) of each modality and perspective for each database.

In Figure 3, if we examine the speech modality in the activation dimension, both databases have their *Main-Perspectives* concentrate on first MFCC with the CIT additionally focuses on intensity, and the NNIME zooms in on pitch. However, each of their *Derived-Perspectives* bring in complementary acoustic descriptors, e.g., the 2nd and 3rd MFCCs with the CIT additionally zooms in on pitch, and the NNIME focuses on intensity. While the exact complementary pattern is more complex in terms of valence dimension, we can still observe the disparate distributions obtained from each of the two emotion perspectives on their selection of important acoustic LLDs.

Furthermore, if we examine video modality, the first thing to note is that the 8th and the 9th bin in HOF carry more weights signifying these movements may be more emotionally-relevant as compared to other bins of movements. The 8th and the 9th bins encode movement toward lower right or lower left in the 2D screen (or simply closer to the camera); this may also capture movements of both speakers coming closer and moving away from each other. If we examine the differences in *Main-Perspective* and *Derived-Perspective* in the two databases, we can see that these movements consistently being integrated from the *Derived-Perspective* but not the *Main-Perspective*. By deriving multiple emotion perspectives, we not only show that they can provide complementary information but also that the additional diverse perspectives can uncover emotionally-relevant behaviors that may be otherwise hard to obtain.

5. Conclusions

In this work, we propose a novel bootstrapped weighted kernel fusion to improve the robustness of our previously approach of cross-corpus emotion recognition. The approach is based on deriving a weighted fused gram matrix, which encodes the importance of the *perspective-dependent* features

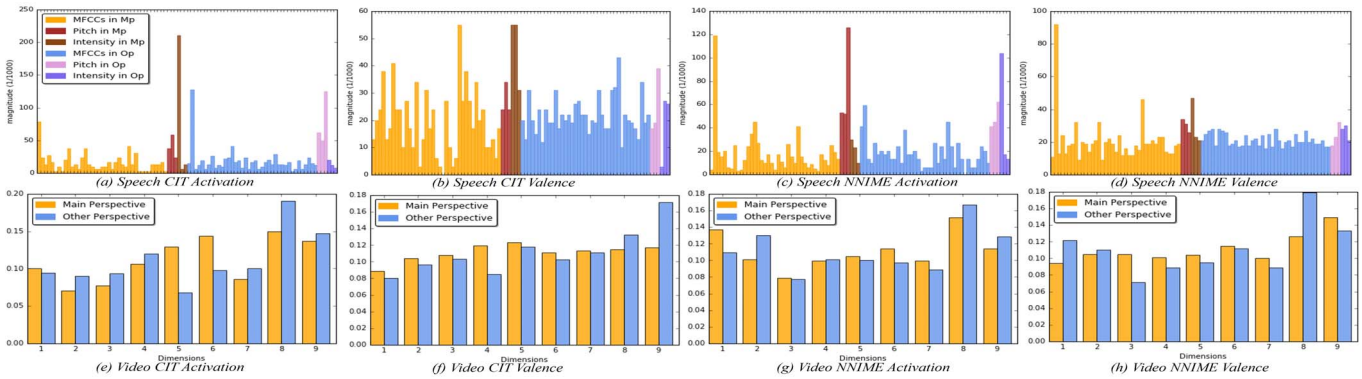

Figure 3. *It shows a plot where x-axis indicates the list of all of the original LLDs and y-axis shows the distribution on the original LLDs out of the top 25% weighted GMM-FV dimensions selected (section 3.3.1) for each modality and perspective in each database.*

when integrating cross-corpus emotion information using kernel-fusion technique. Our approach integrates the joint emotionally-relevant information across both databases. We further bring additional insights on the complementary and advantageous nature of having multiple diverse perspectives in automatic recognition of emotion attributes.

There are several future directions. One of the immediate directions is to extend the framework to include more *factors/settings*, e.g., languages, elicitation styles, cultural backgrounds, etc, to further enrich the modeling power of our multi-perspective fusion architecture. Also, with additional databases and better understanding of each factor, we can develop an informed end-to-end deep learning neural network framework to encapsulate various modules into a single optimization. The continuous advancement toward robust emotion recognizer will provide a key enabler for emerging field of complex human behavioral studies and modeling, e.g., behavioral signal processing (BSP) [18].

Acknowledgments

Thanks to Ministry of Science and Technology (103-2218-E-007-012-MY3).

References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[2] F. Nasoz, K. Alvarez, C. L. Lisetti, and N. Finkelstein, "Emotion recognition from physiological signals using wireless sensors for presence technologies," *Cognition, Technology & Work*, vol. 6, no. 1, pp. 4–14, 2004.

[3] K. Byron, S. Terranova, and S. Nowicki, "Nonverbal emotion recognition and salespersons: Linking ability to perceived and actual success," *Journal of Applied Social Psychology*, vol. 37, no. 11, pp. 2600–2619, 2007.

[4] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and autonomous systems*, vol. 42, no. 3, pp. 143–166, 2003.

[5] Z. Zhang, F. Weninger, M. Wöllmer, and B. 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.

[6] D. Bone, C.-C. Lee, and S. S. Narayanan, "A robust unsupervised arousal rating framework using prosody with cross-corpora evaluation." in *INTERSPEECH*, 2012, pp. 1175–1178.

[7] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[8] B. Zhang, E. M. Provost, and G. 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.

[9] S. M. Feraru, D. 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.

[10] C. M. Chang and C.-C. Lee, "Fusion of multiple emotion perspectives:improving affect recognition through integrating crosslingual emotion information." in *ICASSP*, 2017.

[11] C. Xu, D. Tao, and C. Xu, "Large-margin multi-viewinformation bottleneck," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1559–1572, 2014.

[12] P. Dhillon, D. P. Foster, and L. H. Ungar, "Multi-view learning of word embeddings via cca," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2011, pp. 199–207. [Online]. Available: http://papers.nips.cc/paper/4193-multi-view-learning-of-word-embeddings-via-cca.pdf

[13] S. Sun, "A survey of multi-view machine learning," *Neural Computing and Applications*, vol. 23, no. 7-8, pp. 2031–2038, 2013.

[14] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE transactions on affective computing*, vol. 3, no. 2, pp. 211–223, 2012.

[15] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: face, body gesture, speech," in *Affect and emotion in human-computer interaction*. Springer, 2008, pp. 92–103.

[16] A. Metallinou, Z. Yang, C.-c. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The usc creativeit database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations," *Language resources and evaluation*, vol. 50, no. 3, pp. 497–521, 2016.

[17] H.-C. Chou, W.-C. Lin, L.-C. Chang, C.-C. Li, H.-P. Ma, and C.-C. Lee, "Nnime: The nthu-ntua chinese interactive multimodal emotion corpus," in *Seventh International Conference on Affective Computing and Intelligent Interaction (ACII2017)*. IEEE, 2017.

[18] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.