Empirical Mode Decomposition: A Data-Enrichment Perspective on Speech Emotion Recognition

Bin Dong[1], **Zixing Zhang**[1], **Björn Schuller**[1,2]
[1]Chair of Complex & Intelligent Systems, University of Passau, Passau, Germany
[2]Department of Computing, Imperial College London, London, UK
bin.dong@uni-passau.de, zixing.zhang@uni-passau.de, schuller@IEEE.org

Abstract

To deal with the data scarcity problem for Speech Emotion Recognition, a novel data enrichment perspective is proposed in this paper by applying Empirical Mode Decomposition (EMD) on the existing labelled speech samples. In doing this, each speech sample is decomposed into a set of Intrinsic Mode Functions (IMFs) plus a residue by EMD. After that, we extract features from the primary IMFs of the speech sample. Each single classification model is trained first for the corresponding IMF. Then, all the trained models of the IMFs plus that of the original speech are combined together to classify the emotion by majority vote. Four popular emotional speech corpora and three feature sets are used in an extensive evaluation of the recognition performance of our proposed novel method. The results show that, our method can improve the classification accuracy of the prediction of valence and arousal with different significance levels, as compared to the baseline.

Keywords: Speech emotion recognition, empirical mode decomposition, intrinsic mode function, majority vote, support vector machine

1. Introduction

Speech Emotion Recognition (SER) has attracted increasing interest in the context of speech processing and machine learning (Han et al., 2014), and is going to be implemented in real-life applications like video games (Schuller et al., 2015), health care systems (Tacconi et al., 2008), and service robots (Marchi et al., 2014). One bottleneck of these applications, however, is the scarcity of labelled data that are yet necessary to build robust machine learning systems (Sainath et al., 2015).

To overcome the problem of data scarcity for SER, some studies have been done in the past few years. The work in (Schuller et al., 2011) attempted to make efficient use of multiple available small size of annotated databases to develop a robust model by the strategy of pooling or voting. Nevertheless, the majority of speech emotional databases that are publicly available at present have only a few hours of annotated instances (Schuller et al., 2010). In contrast to these limited labelled data, unlabelled data seem countless and can be easily collected. To exploit the large amount of unlabelled data, the approach of Semi-Supervised Learning (SSL) (Zhang et al., 2011) and its advanced derivations like Co-Training (Liu et al., 2007) were proposed and investigated, and showed much better performance than the approach which merely uses labelled data. Later on, Active Learning algorithms by sparse instance tracking (Zhang and Schuller, 2012) and label uncertainty (Zhang et al., 2015) were studied with aim at achieving higher accuracy with less human work for labelling the selected samples.

To further deal with this data scarcity problem, the present paper proposes a novel prospective to utmost exploit the existing labelled speech samples. It uses Empirical Mode Decomposition (EMD) to decompose the original speech sample into a set of Intrinsic Mode Functions (IMFs), each of which can be regarded as a specific counterpart of the original speech sample in a limited frequency band(Huang et al., 1998), which could provide additional information for the systems. Inspired by the idea of P. Flandrin *et al* – EMD wokrs as a filter bank (Flandrin et al., 2004), we can consider EMD as the operation which decomposes the nonlinear and nonstationary speech sample into the quasi-linear and quasi-stationary components – IMFs. In doing so, the number of speech samples will be multiple-fold increased.

In the following, we investigated the proposed data enrichment method for SER in terms of three items: 1) decompose each original speech into a set of IMFs (plus a residue); 2) extract three popular feature sets not only on the original speech sample but also on its primary IMFs; 3) apply to four widely used speech emotional corpora (spontaneous and unspontaneous).

The remainder of the paper is organized as follows. Section 2 introduces the method of EMD for enriching the speech samples and the following emotion recognition based on the enriched samples. The performance of the proposed method is evaluated by three feature sets and four popular emotional corpora and then compared with baseline results in Section 3. Based on the recognition results, we discuss the performance of our method and make conclusions at the end of Section 4.

2. Empirical Mode Decomposition for Data Enrichment

Since the voiced part of the speech is more important to analyse emotion and to save computation, only the voiced parts of the recordings are decomposed by EMD in the present paper. Furthermore, the decomposition speed of EMD strongly depends on the length of the sample. The sum of the time of decomposing each single voiced part is much less than the time of decomposing the sum of all voiced parts.

2.1. Localization of Voiced Parts

To detect and locate the voiced parts in a speech sample, one method, named YAAPT (Yet Another Algorithm for Pitch Tracking) (Zahorian and Hu, 2008), is applied. It was originally issued to robustly track the fundamental frequency F_0 of the target speech. We can use the results of YAAPT to determine the positions and durations of the voiced parts in the speech.

A discrete speech sample is denoted as $x(n)$ with $n = 1, 2, \cdots, N$. Without loss of generality, the algorithm YAAPT can be treated as an abstract function $f\{\cdot\}$ which maps the speech $x(n)$ to its fundamental frequency F_0:

$$F_0(m) = f\{x(n)\}, \tag{1}$$

where $m = 1, 2, \cdots, M$ and the relationship between M and N depends on the length of and the overlapping of the sliding window in YAAPT. Then the nonzero elements in $F_0(m)$ are mandatorily set to 1 and the normalized $F_0(m)$ is written as $\hat{F}_0(m)$ which consists of 0 and 1 only. Then it calculates the finite difference of $\hat{F}_0(m)$, $\Delta\hat{F}_0(m)$ with just three values -1, 0, and 1. In the value set of $\Delta\hat{F}_0(m)$, most elements are 0 and only a few ones are -1 and 1, which are in pairs. The value 1 indicates the starting of one voiced part and the following -1 as its ending. Therefore, once the indices of the elements 1 and -1 are fully determined, the starting and ending indices of all the voiced parts will be easily calculated by using Eq.(1) for the original speech. After that, the speech is segmented based on these voiced infromation. Due to the space limit, the algorithm YAAPT is not introduced here in detail.

2.2. Data Enriching by EMD

After detecting the voiced parts of recording, an derived EMD algorithm called CEEMDAN (Complete Ensemble Empirical Mode Decomposition with Adaptive Noise) is applied to decompose the truncated voiced parts. The advantage of CEEMDAN is that it not only can effectively remove the mode mixing from IMFs, but also provides less IMFs than EMD, which will save more calculation for the following feature extraction and emotion classification. For the details of CEEMDAN, the readers are invited to refer to the paper (Torres et al., 2011).

Here is the truncation of the i-th voiced part $x_i(l)$ with $i = 1, 2, \cdots, N_i$. N_i denotes the total number of the voiced parts in the speech sample. After executing CEEMDAN, we can rewrite $x_i(l)$ as

$$x_i(l) = \sum_{k=1}^{K} c_{[i,k]}(l) + r_i(l), \tag{2}$$

where $c_{[i,k]}(l)$ stands for the k-th IMF of the i-th voiced part $x_i(l)$, K for the total number of the IMFs, and $r_i(l)$ for the decomposition residue of $x_i(l)$.

The characteristic frequencies of IMFs decrease with the increasing of their indices k. For emotion recognition, the IMFs whose characteristic frequencies are lower than the fundamental frequency F_0 are useless. They occupy very little proportion of the energy of the original speech. Moreover, they are inaudible to us, no matter how large amplification coefficients are applied. To save cost, these trivial IMFs are deliberately ignored from now on.

If the sampling frequency of the original speech is very high, for instance 44.1 kHz, its first several IMFs also need to be discarded. The IMFs whose characteristic frequencies are higher than 10 kHz, act as noise and can not provide useful information for the following emotion classification. When we extract their features in terms of the feature set – eGeMAPS (Eyben et al., 2016) for example, most features could not get valid values. Therefore, only the middle IMFs are kept as the primary ones for the following feature extraction and emotion classification. In current stage, the selection of the primary IMFs depends on their energy and audio content. When the sampling frequency is 16 kHz or less, the selection of the primary IMFs starts from IMF 1. Note that the number of the primary IMFs is suggested to be odd for the benefit of the following majority vote.

After determining the primary IMFs of all voiced parts, we combine them together to generate the IMFs of the original speech in terms of the sequence of the voiced parts. For instance, the k-th IMF $c_k(n)$ of the speech sample $x(n)$ can be represented by $\mathbf{c}_{[i,k]}$ as $\mathbf{c}_k = [\mathbf{0}, \mathbf{c}_{[1,k]}, \mathbf{0}, \mathbf{c}_{[2,k]}, \mathbf{0}, \cdots, \mathbf{c}_{[N_i,k]}, \mathbf{0}]$, where $\mathbf{c}_k$ is the vector denotation of $c_k(n)$ and the vector $\mathbf{0}$ replaces the corresponding unvoiced part in the original speech. Then the following feature extraction will be directly conducted on the reconstructed IMFs $\mathbf{c}_k$ one by one.

2.3. Speech Emotion Recognition

After extracting the features from the speech samples and their primary IMFs by using openSMILE (Eyben et al., 2010), we begin to train classification models for the original speeches and their IMFs one by one. That means each primary IMF only employs its own features to train a specific model, but does not employ the features of the other IMFs and not share a common model with the other IMFs. We apply a classic algorithm – Support Vector Machine (SVM) to execute the emotion classification for each single IMF. The whole classification can be represented as follows

$$\mathcal{H}(\mathbf{v}) = \arg\max_{y \in \mathcal{Y}} (w \cdot 1(y = h(\mathbf{v}) + \sum_{i=1}^{R} 1(y = h_i(\mathbf{v}_{IMF}))), \tag{3}$$

where $\mathbf{v}$ and $\mathbf{v}_{IMF}$ are the feature vectors of the original speech and of its primary IMFs, respectively; the symbol $\mathcal{Y}$ denotes a prediction space; the value of $1(a)$ is 1 if a is true and 0 otherwise; w represents the weight of the original speech sample; and R is the number of the primary IMFs. Note that the primary IMFs of the speech sample are treated equally in the majority vote and their weighting coefficients are all set to 1.

Although the original speeches can provide much information as references for emotion recognition, nobody knows how much useful information the original samples can provide, comparing with that of their IMFs, to the majority vote of the final emotion classification. To investigate the significance of the original speeches on the majority vote, we employ three different weighting coefficients ($w = 0, 1, 2$) here. In detail, $w = 0$ means that no origi-

nal speech takes part in the majority vote; $w = 1$ suggests that the original speeches are treated the same as their own IMFs in the majority vote; $w = 2$ signifies that the original speeches are considered to be more important than their own IMFs in the majority vote. It can be replaced by other coefficients, such as $w = 3$ and $w = 4$, but their final emotion recognitions improve little.

3. Empirical Experiments

In this Section, we focus on the performance evaluation of the method introduced in Section 2.

3.1. Emotional Corpora and Feature Sets

To comprehensively evaluate our method, four emotional corpora were applied here: Geneva Multimodal Emotion Portrayals (GEMEP) corpus (Bänziger et al., 2012), the eNTERFACE05 Audio-Visual Emotion Database (Martin et al., 2006), the Vera am Mittag (VAM) German audio-visual emotional speech database (Grimm et al., 2008), and FAU AIBO corpus (Batliner et al., 2008). The first two corpora (GEMEP and eNTERFACE) are unspontaneous, but the last two (VAM and FAU AIBO) are spontaneous. Their introduction is listed in Table 1.

The selected feature sets are the popular ones: a) *eGeMAPS* (Eyben et al., 2016) with 88 highly efficient feature, b) *InterSp09* (Schuller et al., 2009) with 384 parameters used for the 2009 INTERSPEECH Emotion Challenge, and c) *InterSp13* (Schuller et al., 2013) with 6373 parameters consequently employed for the 2013-2015 INTERSPEECH Paralinguistics Challenges. To extract these features, the toolkit of openSMILE (Eyben et al., 2010) was implemented in the present investigation.

3.2. Performance Evaluation

As to the classifier, we use the standard SVM initially trained with a Sequential Minimal Optimization (SMO) algorithm with a linear kernel and a complexity constant of 0.05. In terms of performance evaluation, we use the Unweighted Average Recall (UAR).

Table 2 shows the SER performance of our approach based on data enrichment. From the table, we find three points:

1) Generally speaking, our proposed approach could deliver better results in comparison with the baseline by using three different feature sets. Particularly, *InterSp13* shows the best results not only of the baseline, but also of the improvement of our approach.

2) The proposed approach for the task of valence performs better than that of arousal: 10 wins out of 12 cases for valence *vs.* 5 wins out of 9 cases for arousal. Particularly, the performance improvement on the database of FAU AIBO shows an significance level at 0.01 when using *InterSp13* in all three cases ($w = 0, 1, 2$).

3) The weighting coefficients apparently affect the emotion recognition. The weighting coefficient $w = 2$ works best and it offers two significant improvements: the valence on FAU AIBO and eNTERFACE. Instead, without taking into account the original speeches (i.e. the weighting coefficient $w = 0$), the accuracy of the emotion recognition is improved trivially, in addition to the valence on FAU AIBO with *InterSp13*.

4. Discussion and Conclusions

Based on the findings in Subsection 3.2, we further discuss the results of the proposed method.

As pointed by M. Goudbeek and K. Scherer, the arousal mainly depends on the fundamental frequency F_0 and intensity measures, but the valence are related to duration and spectral balance (i.e. spectral shape parameters) (Goudbeek and Scherer, 2010). The function of EMD is to decompose a signal into a set of analytical components – IMFs. After analysing the characters of all IMFs in the time and frequency domains and listening to their audio contents, we find that only one IMF strongly correlates the fundamental frequency F_0 of the original speech sample, for example, IMF 10 at the sampling frequency 44.1 kHz and IMF 6 at 16 kHz. Thus the other primary IMFs can not provide valid values for the parameter F_0 and the provided values are usually the integer times of the fundamental frequency of the original speech.

The intensity of the original speech is equal to the sum of the intensity of all IMFs and the residue in terms of the superposition principle. When the intensity measures of the primary IMFs are calculated, their values are less than that of the original speech. Therefore, the proposed method does not work quite well on the recognition of arousal, although the majority vote can correct the distortion in the feature values partly.

Instead, no matter how EMD is executed, the durations of utterances in the selected IMFs basically keep the same as those in their original speech samples. As pointed out by P. Flandrin *et al.* (Flandrin et al., 2004), EMD works as a filter bank. The spectral shape of each primary IMF keeps similar with that of the original speech in the same frequency band where the IMF stays. Furthermore, each selected IMF can provide more details of spectral shape in their working frequency band, i.e. more bins in the specific frequency band. That means the primary IMFs provide more accurate spectral information for the following valence recognition. That is the reason why the proposed methods outperforms the baseline on the valence recognition, especially the valence recognition on the corpus FAU AIBO with the feature set *InterSp13*.

Second, involving the original speech samples is significant for the majority voting. As the applied feature sets are not fully compatible to the primary IMFs, we need to take their original speeches as the references. To highlight the significance of the references, it is better to endow more weighting to the original speeches. Obviously, it is only the temporary way to combine the original speeches into the majority vote for the emotion recognition, as we have not found the most suitable feature set for the proposed method until now. In future, we only apply the features extracted from the primary IMFs to recognize emotions, but without any features of their original speech.

The idea feature set for our method should meet at least two criteria: 1) the features extracted from the primary IMFs are enough for the following emotion recognition; 2) the recognition accuracy of our method with the idea feature

Corpus	Lan	Emotion	# Arousal		# Valence		# All	# m	# f	Rec
			-	+	-	+				
GEMEP	Fr	acted	2,520	2,520	2,520	2,520	5,040	5	5	studio
eNTERFACE	En	induced	425	852	855	422	1277	34	8	studio
VAM	De	natural	501	445	875	71	946	15	32	noisy
FAU AIBO	De	induced			5,823	12,393	18,216	21	30	studio

Table 1: Overview of the selected emotion corpora (Lan: language, Rec: recording environment, f/m: (fe-)male subjects).

Corpus	Arousal [%]				Valence [%]			
	Base	$w = 0$	$w = 1$	$w = 2$	Base	$w = 0$	$w = 1$	$w = 2$
(a) eGeMAPS								
GEMEP	79.0	**80.3**	80.1	79.9	61.3	61.3	61.6	**61.8**
eNTERFACE	**71.5**	68.2	68.3	68.3	64.2	66.8	**68.3**	**69.7***
VAM	**79.5**	73.0	74.2	75.4	48.4	**50.6**	46.0	41.4
FAU AIBO					**68.3**	67.3	67.5	67.8
(b) InterSp09								
GEMEP	**82.4**	81.0	81.4	81.7	63.0	64.1	65.3	**66.4**
eNTERFACE	73.7	71.8	73.5	**75.2**	**71.0**	63.3	65.9	68.5
VAM	68.6	**72.5**	72.2	71.9	40.1	**46.6**	44.9	43.2
FAU AIBO					68.5	68.3	68.6	**68.8**
(c) InterSp13								
GEMEP	79.2	80.8	81.8	**82.9**	66.2	66.2	66.9	**67.6**
eNTERFACE	**80.0**	69.0	72.7	76.3	75.0	70.2	73.3	**76.5**
VAM	75.6	**80.4**	79.9	79.3	47.5	**51.2**	50.4	48.4
FAU AIBO					65.4	**68.0****	**67.6****	**67.3****

Table 2: The Unweighted Average Recall (UAR) of proposed data enrichment approach on four emotional corpora (GEMEP, eNTERFACE, VAM, and FAU AIBO) by the combination of Empirical Mode Decomposition and Majority Vote with different weighting coefficients defined on the original speech sample: $w = 0, 1, 2$. The feature sets are (a) the extended Geneva Minimalistic Acoustic Parameter Set (*eGeMAPS*), (b) the INTERSPEECH 2009 Emotion Challenge (*InterSp09*), and (c) the INTERSPEECH 2013 Computational Paralinguistics Challenge (*InterSp13*). The corresponding baseline results are also listed here. The symbols of * and ** denote the significant levels of performance improvement at 0.05 and 0.01 by one-tailed hypothesis, respectively.

set outperforms (or at least is comparable with) that of the currently popular methods, for example those listed in Ref. (Eyben et al., 2016).

At last, short conclusions are made here. We proposed a data enrichment approach by using EMD to decompose each original speech sample into a set of IMFs plus a residue, which can serve as additional speech samples to enlarge the size of training sets. Four databases with a variety of languages and speech styles, and three popular feature sets are implemented to evaluate the performance of the proposed approach. The experiments show that method can remarkably increase the recognition accuracy of emotion speeches. It works well not only with the nonspontaneous emotional corpora (GEMEP and eNTERFACE), but also with the spontaneous ones (VAM and FAU AIBO). Future work will exploit new feature set to fit our decomposed samples – the primary IMFs.

5. Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme through the ERC Starting Grant No. 338164 (iHEARu), and the European Union's Horizon 2020 Programme through the Research Innovation Actions No. 645094 (SEWA), No. 644632 (MixedEmotions), and No. 645378 (ARIA-VALUSPA), and by the German Federal Ministry of Education, Science, Research and Technology (BMBF) under grant agreement #16SV7213 (EmotAsS).

Bänziger, T., Mortillaro, M., and Scherer, K. R. (2012). Introducing the geneva multimodal expression corpus for experimental research on emotion perception. *Emotion*, 12(5):1161–1179.

Batliner, A., Steidl, S., and Nöth, E. (2008). Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus. In *Proc. of a Satellite Workshop of LREC 2008 on Corpora for Research on Emotion and Affect*, pages 28–31, Marrakech, Morocco.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). openSMILE – the Munich versatile and fast open-source audio feature extractor. In *Proc. of ACM MM*, pages 1459–1462, Florence, Italy.

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., Andre, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., and Truong, K. (2016). The geneva min-

imalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*. to appear.

Flandrin, P., Rilling, G., and Goncalves, P. (2004). Empirical mode decomposition as a filter bank. *Signal Processing Letters, IEEE*, 11(2):112–114, Feb.

Goudbeek, M. and Scherer, K. (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America*, 128(3):1322–1336.

Grimm, M., Kroschel, K., and Narayanan, S. (2008). The Vera Am Mittag German audio-visual emotional speech database. In *Proc. of ICME*, pages 865–868, Monterrey, Mexico.

Han, K., Yu, D., and Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Proc. of INTERSPEECH*, pages 223–227, MAX Atria, Singapore.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N. C., Tung, C. C., and Liu, H. H. (1998). The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1971):903–995.

Liu, J., Chen, C., Bu, J., You, M., and Tao, J. (2007). Speech emotion recognition using an enhanced co-training algorithm. In *Proc. of ICME*, pages 999–1002, Beijing, China.

Marchi, E., Ringeval, F., and Schuller, B. (2014). Voice-enabled assistive robots for handling autism spectrum conditions: An examination of the role of prosody. In A. Neustein, editor, *Speech and Automata in the Health Care*, pages 207–236. Walter de Gruyter GmbH & Co KG.

Martin, O., Kotsia, I., Macq, B., and Pitas, I. (2006). The eNTERFACE'05 audio-visual emotion database. In *IEEE Workshop on Multimedia Database Management*, pages 8–15, Atlanta, GA.

Sainath, T. N., Weiss, R. J., Senior, A., Wilson, K. W., and Vinyals, O. (2015). Learning the speech front-end with raw waveform CLDNNs. In *Proc. of INTERSPEECH*, pages 1–5, Dresden, Germany.

Schuller, B., Steidl, S., and Batliner, A. (2009). The INTERSPEECH 2009 emotion challenge. In *Proc. of INTERSPEECH*, pages 312–315, Brighton, UK.

Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., and Rigoll, G. (2010). Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131.

Schuller, B., Zhang, Z., Weninger, F., and Rigoll, G. (2011). Using multiple databases for training in emotion recognition: To unite or to vote? In *Proc. of INTERSPEECH*, pages 1553–1556, Florence, Italy.

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., and Kim, S. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proc. of INTERSPEECH*, Lyon, France. no pagispanation.

Schuller, B., Marchi, E., Baron-Cohen, S., Lassalle, A., O'Reilly, H., et al. (2015). Recent developments and results of asc-inclusion: An integrated internet-based environment for social inclusion of children with autism spectrum conditions. In *Proc. of IDGEI*, Atlanta, GA. no pagination.

Tacconi, D., Mayora, O., Lukowicz, P., Arnrich, B., Setz, C., Troster, G., and Haring, C. (2008). Activity and emotion recognition to support early diagnosis of psychiatric diseases. In *Proc. of PervasiveHealth*, pages 100–102, Istanbul, Turkey.

Torres, M. E., Colominas, M. A., Schlotthauer, G., and Flandrin, P. (2011). A complete ensemble empirical mode decomposition with adaptive noise. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4144–4147, May.

Zahorian, S. A. and Hu, H. (2008). A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America*, 123(6):4559–4571.

Zhang, Z. and Schuller, B. (2012). Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition. In *Proc. of INTERSPEECH*, Portland, OR.

Zhang, Z., Weninger, F., Wöllmer, M., and Schuller, B. (2011). Unsupervised learning in cross-corpus acoustic emotion recognition. In *Proc. of IEEE workshop on Automatic Speech Recognition and Understanding*, pages 523–528, Big Island, HY.

Zhang, Y., Coutinho, E., Zhang, Z., Quan, C., and Schuller, B. (2015). Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions. In *Proc. of ICMI*, pages 275–278, Seattle, WA.