ON THE MODELING OF NATURAL VOCAL EMOTION EXPRESSIONS THROUGH BINARY KEY

Jordi Luque, Xavier Anguera

Telefonica Research
Edificio Telefonica-Diagonal 00, Barcelona, Spain
{jls, xanguera}@tid.es

ABSTRACT

This work presents a novel method to estimate natural expressed emotions in speech through binary acoustic modeling. Standard acoustic features are mapped to a binary value representation and a support vector regression model is used to correlate them with the three-continuous emotional dimensions. Three different sets of speech features, two based on spectral parameters and one on prosody are compared on the VAM corpus, a set of spontaneous dialogues from a German TV talk-show. The regression analysis, in terms of correlation coefficient and mean absolute error, show that the binary key modeling is able to successfully capture speaker emotion characteristics. The proposed algorithm obtains comparable results to those reported on the literature while it relies on a much smaller set of acoustic descriptors. Furthermore, we also report on preliminary results based on the combination of the binary models, which brings further performance improvements.

Index Terms— Emotion modeling, binary fingerprint, VAM corpus, dimensional emotions

1. INTRODUCTION

Emotional expression is a natural modulator of human interactions in speech communications. Empathy in a conversation and effective social interaction depends upon the ability to accurately perceive and express emotions. Most of the current work in emotion recognition assumes a categorical representation of emotions [1, 2], formulated as a classification task. This is not in alignment with the psychology theory that claims for the importance of a continuous emotional space analysis [3]. Nonetheless, acoustic analysis of the signal must take into account models defining the sentiments being measured in a continuous way rather than mapped to discrete categories [4]. To our knowledge, the first study to directly correlate speech with the continuous emotion dimensions is presented in [5, 6]. They employed a set of features derived from the pitch and the energy contour of the speech waveform and others related to speaking rate and spectral characteristics of the signal, leading to a total of 46 features, which were then modeled using a Support Vector Regression (SVR) model. A similar experiment but with a different set of features was also reported in [7].

Up to now, among the different works present in the literature, there is no a common agreement on the acoustic feature sets and modeling techniques most suitable for the representation and modeling of human emotions. Such a situation is likely the consequence of the limited amount of emotion-tagged data and the variety of speech databases employed for characterizing emotions which do not account for speaker and session variability [1, 8]. Nevertheless, some configurations are more preferred than others. On the modeling side, the use of Support Vector Machine (SVM) classifiers with radial basic function seems to outperform other techniques [1,6] due to their robustness to over-fitting and their discriminative power. On the feature selection side, recent comparisons at Interspeech challenges [9] have proven that MFCC are among the most suitable descriptors for recognizing emotions in speech. In [8, 10] extensive studies of emotional speech data and the effects on the selection of the set of features are reported.

In this paper we evaluate the use of an SVR modeling approach in the binary domain by first "binarizing" the input features to obtain a single binary fingerprint per input utterance. The binarization of the feature space, initially proposed in [11] for speaker recognition, has been successfully applied to categorical emotion recognition in [12] and is the first time that it is applied for continuous emotion recognition. To test the proposed system we experiment with three different sets of acoustic features. These are the standard spectral low-MFCC features, source excitation features extracted from Linear Prediction Coding (LPC) residual signal and prosodic features derived from pitch. We show that even though such features are of much lower dimensionality than standard feature sets used recently for the task [13] and only represent a very local time-portion of the signal, by using binarization we are able to indirectly incorporate long term information and successfully model emotions. Furthermore, given that the emotion fingerprints are represented by a binary vector it is much simpler to interpret how an emotion is represented (and differs from others) and see the differences between different emotions.

Experimental results on the VAM corpus [14] show that

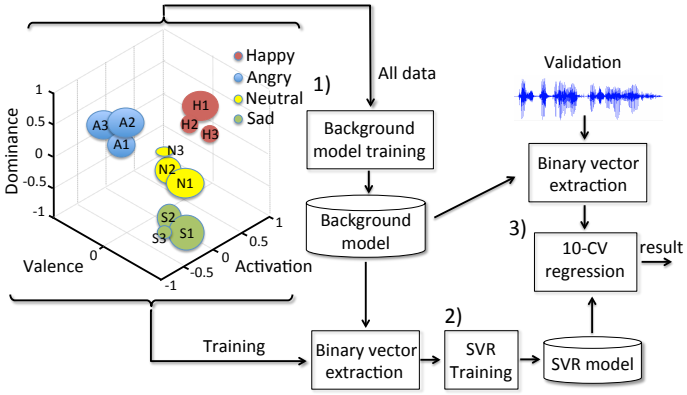

Fig. 1: *Binary emotion regression system, comprised of three main modules: 1) Background model training; 2) emotion regression training; 3) cross-validation*

our approach achieves comparable results to those reported in the literature [7, 14], but using a frame-based feature approach and a novel binary modeling technique.

2. EMOTION MODELING

The proposed approach for continuous modeling of emotions (also referred as emotion regression) builds upon the acoustic binary fingerprints initially proposed in [11] for speaker recognition and adapted in [12] for emotion classification. Figure 1 shows the main modules involved in the emotion regression system, both for training and for validation.

In the training phase a set of acoustic utterances needs to be provided together with their perceived emotional values. In this paper we use the VAM corpus [14], which consists of training utterances labelled by 6 independent labelers in terms of Valence, Activation and Dominance. Each utterance is first processed to obtain a set of acoustic features, as explained in section 2.1. Then, a single binary fingerprint is computed from the whole utterance by using one or several background models, as described in Section 2.2. With these binary samples we train an SVR model [15] to estimate a regression for each of the emotional dimensions. In the validation phase, test utterances are first mapped into a binary fingerprint in the same way as we map training utterances. Then, using the SVR model, the output emotional values are guessed. Following we give a detailed description of each of these steps.

2.1. Acoustic Feature Extraction

Three different sets [1] of acoustic features have been used in this work:

1. A set of 12 low-band MFFC (ranging from 20Hz to 350Hz) are extracted from 24 Mel-scaled logarithmic filters and augmented with log-energy. In addition, the derivative and acceleration of all features is taken, obtaining a total of 39 MFCC coefficients. These are extracted every 10ms with a standard analysis window of 25ms. We expect these MFCCs to model F0 variations, as localized correlation exists between fundamental frequency and spectral envelope [16].
2. Standard MFCC features (same configuration as above, but for the full signal bandwidth) are obtained from the residual signal resulting from the Linear Prediction modeling (using 12 LPC coefficients) of the original speech wave. We expect them to model non-linear F0 variations and contribute with an extra amount of formant information modeling that is not present in conventional MFCC coefficients.
3. A short-time pitch estimation based on an autocorrelation method on the voiced regions of speech [17]. In addition, we also extract the log-energy, the derivative and acceleration of the pitch feature, resulting in a six dimensional feature vector. We expect these to capture traits of the acoustic realization of prosody by variations of the fundamental frequency. These features are standardized through mean and standard deviation estimated per speaker.

Low-band MFCC features have been reported to perform better in emotion recognition tasks than wide-band MFCC or pitch related features [18], whereas short-time suprasegmental features are believed to highly correlate with emotions. For this reason we decided to extract the three feature sets and to compare their performance when obtaining their binary fingerprints.

Given an input training or test utterance, either one of the feature sets is extracted and then converted into a single binary fingerprint, as explained in Section 2.2. In the experimental section we compare the performance of each set for different background model sizes.

2.2. Fingerprints Extraction

The process which maps a set of M acoustic feature vectors, extracted from an input utterance, into a single binary vector [11] involves the training of an acoustic background model. This model has been previously estimated by means of unlabeled data similar to that we later use in the test via unsupervised EM-ML training. The number of Gaussian mixtures N in this model define the dimension of the output binary vector (fingerprint). Intuitively, the active bits in the fingerprint indicate where in the acoustic space the acoustic data is mostly found. The proposed method shares some details with the UBM weight posterior probability (UWPP) method proposed in [19, 20]. In such method, occupancy posterior probabilities per Gaussian are stacked in a supervector which feeds a classifier. In our approach we just select the most informative posteriors in order to map them to value 1 in a binary vector. Through experimentation we have seen our method to be less affected by noise.

In order to obtain the binary fingerprint we first convert all acoustic feature vectors $x_i, i = 1 \dots M$, representing a given emotion value (see fig. 3), into binary form. It is done

[1] MFCC and pitch related features has been computed using HTK and Praat software packages

by selecting the 5 Gaussians[2] in the background model with highest posterior probability ($P(x_i|\lambda_j), j = 1 \dots N$), where j stands for the Gaussian number in the UBM model. Next we combine the information obtained by each individual feature vector into a single binary fingerprint. This is done with a vector of counts v_c where each position counts how many times each Gaussian in the background model has been selected by any of the individual feature vectors. Note that v_c contains a total number of counts equal to 5 times M, containing a histogram of which Gaussians are most relevant to model the input data. Note also that each input feature vector casts the same number of *votes* independently from the other features. This is useful to cope with heterogeneous information usually present in the speech signal (e.g. small silence regions or acoustic artifacts) other than the speech that we desire to model. Finally, we convert the vector of counts v_c into binary form by choosing the 25%-highest count dimensions to become 1, setting to 0 the rest. This threshold was empirically chosen as higher values resulted in lower regression performances.

Through this binary fingerprint we are averaging the information obtained at feature level regarding where in the acoustic space the input features mostly occur. This can be considered equivalent to the averaging performed by state-of-the-art emotion recognition systems. These systems represent the whole utterance using several functionals computed from the set of acoustic features extracted from the signal, like in our case, at short-term intervals. The main difference here is that each bit in the fingerprint directly represents the activation status of a given Gaussian in the background model, where each Gaussian represents a region in the acoustic space. We can therefore directly obtain a relationship of which acoustic regions are most prevalent for each one of the emotions, being able to easily interpret results.

In addition to building a binary fingerprint from the input acoustic features and a given background model, we have also experimented with stacking the fingerprints from background models of different sizes. As we will show in the experimental section, the sampling of the acoustic space at different resolutions is able to further improve the final results.

2.3. Regression Modeling and Estimation

As explained in Section 2.2, a fingerprint vector is computed for each set of observations representing an emotional state value. Therefore, each value in the emotional dimension, if exists in the training set, is encoded by means of a binary vector. To investigate whether the fingerprint modeling stands for a good representation of emotions into the 3-D representation model, we train regression models that map the fingerprint features to their corresponding emotion dimension value. The figure 3 depicts the histogram of values per each emotional dimension.

[2]We have chosen to select 5-best after experimenting with different number of selected Gaussians

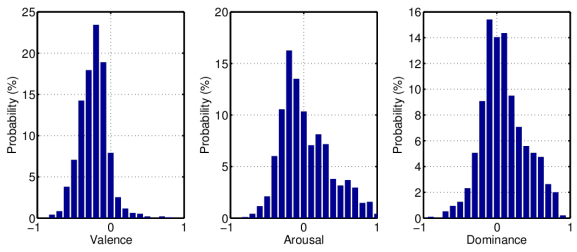

Fig. 3: *From left to rigth, histogram of the emotional dimension values in the VAM corpus per valence, arousal and dominance dimensions respectively.*

The goal here is to estimate a function $f\colon \mathbb{B}^N \rightarrow \mathbb{R}$, where $\mathbb{B}$ represents the whole set of binary vectors and N stands for the complexity of the background model, that is, the dimension of the fingerprint vector. In order to estimate such a function we employed support vector regression. Concretely, nonlinear ϵ-regression that, unlike least square regression, defines the error function to minimize as a ϵ-insensitive loss function [15]. The kernel employed is a radial basis function. Note that the main difference in our implementation compared to that of [6, 7] are both the high number of heterogeneous features employed and that the SVR models binary fingerprints instead of low level descriptors or functionals. It is worth to note that in [7] the authors used a polynomial kernel function of degree 1. As can be seen in the experimental section, the fingerprint vector, which summarizes the statistics of the original frame-based data into a single binary representation, highly correlates with the values in the emotional 3-D model.

Experiments are performed using a 10-fold cross validation (CV) scheme, as in [6, 7], for each feature set independently but selecting folds randomly. It is also worth to mention that a grid search around standard values is performed to estimate C and σ parameters using a different 10-fold cross validation where folds are also randomly selected. Once C and σ are fixed, the regression coefficient ρ is mean-averaged among validation folds in the final 10-fold CV stage.

3. EXPERIMENTAL SECTION

3.1. Database and metrics

In order to evaluate the proposed system we used the VAM corpus [14]. It consists of 12 hours of audio-visual recordings of the German TV talk show "Vera am Mittag". The audio stream, originally sampled at 44.1 kHz was downsapled at 16 kHz previous to the feature extraction step. The database contains 47 speakers, 36 female voices and 11 male voices segmented into broadcasts, dialogue acts and utterances. A total of 1002 different sentences were evaluated by human listeners[3] using self assessment manikins methodology [5] in order to obtain the values in the continuous emotional dimensions, that is, in terms of valence, activation and dominance.

[3]VAM corpus was labelled by 6 evaluators and the inter-evaluator agreement was measured by determining the standard deviation and correlation coefficient among them. See [5].

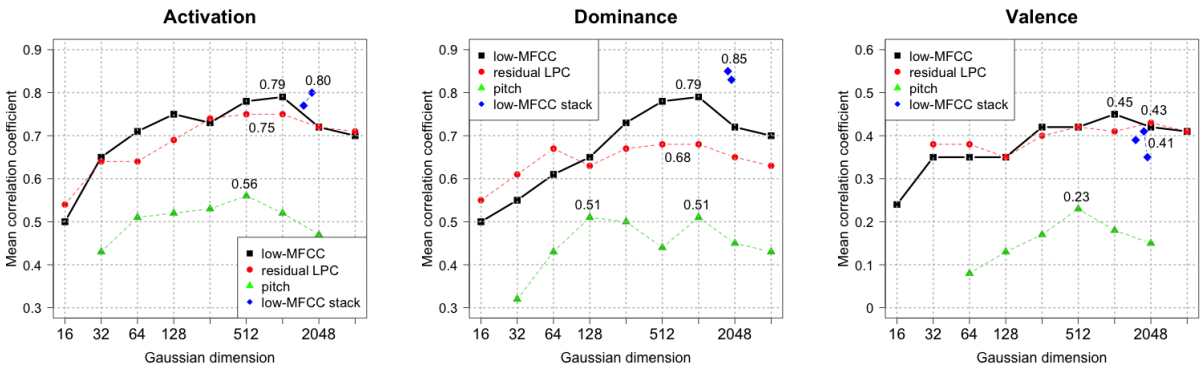

Fig. 2: *Mean correlation coefficient (%) per activation/arousal (left), dominance (middle) and valence (right) depending on the complexity of the backgorund model. Result curves are drawn for the different set of acoustic features: low-MFCC (black squared-points), residual LPC (red circle-points), pitch related (green triangle-points), stacked low-MFCC (blue diamond-points)*

Table 1: *Mean correlation coefficient and mean regression error (in brackets) of the fingerprint-SVR regression estimation on the VAM corpus. First and second rows correspond to the results reported by [6, 7]. Next rows summarize the results reached by the best background model size depending on mean correlation curves in figure 2.*

	Activation	Dominance	Valence
M. Grimm et al. [6]	0.82 (0.15)	0.79 (0.14)	0.46 (0.13)
F. Eyben et al. [7]	0.83 (0.15)	– (–)	0.42 (0.14)
low-MFCC stacked	0.80 (0.16)	0.85 (0.13)	0.41 (0.13)
low-MFCC	0.79 (0.18)	0.75 (0.17)	0.45 (0.14)
LPC residual	0.75 (0.19)	0.68 (0.19)	0.42 (0.14)
pitch+E+A+D	0.56 (0.24)	0.51 (0.22)	0.23 (0.16)

The performance of the binary fingerprint approach is assessed through the mean estimation error and the mean correlation coefficient, computed as the average of each validation fold in cross-validation. Both measures offer us an indication of the predictive ability for the fingerprint-SVR model.

3.2. Experiments

The figures in 2 show the mean correlation coefficient obtained in cross-validation experiments for the three emotion dimensions. Different results are depicted depending on both the complexity of the background model (i.e. the number of Gaussians), and the feature set employed to parametrize the speech data.

A summary of the results for the best setting and for the different feature sets is summarized in table 1. Similar error values to the ones reported in [6] are obtained by the proposed binary approach, with slightly lower correlation coefficients and mean regression errors. In some cases, as in dominance dimension, results reported outperform those in [6]. Also in the dominance dimension the binary fingerprint estimated from low-MFCC features clearly outperforms the other two sets of LPC and pitch based features. Nonetheless, the correlate values of LPC-residual features are close to those from the low-MFCC for the activation and valence dimensions. In general, we observe a poor performance of the pitch related feature set in the whole set of the experiments.

The best results, in terms of correlation and regression error, are reached by the combination of low-MFCCs features coming from three individual fingerprints. In the low-MFCC stacked system the regression analysis is performed on the combined fingerprint vector. The best model error and correlation values for this approach are reported in table 1 and correspond to blue-diamond points in the figure 2. It is a stacked low-MFCC fingerprint vector with a total of 1792 binary values corresponding to aggregate the fingerprints of sizes $1024, 512$ and 256. In addition, 2-best combinations are also drawn in the figures. Over all, the mean regression error lies in between 0.13 and 0.16 and the mean-averaged correlation for the three dimensions is 0.69, similar values than those reported in [6].

3.3. Discussion

From table 1 we observe that the lowest correlation coefficient corresponds to the valence dimension, although the associated mean regression error is lower than that for the other dimensions. It is due to the range of differences in the valence dimension, which presents a very narrow histogram in comparison to activation and dominance histograms, as shown in figure 3. The performance of the fingerprint for MFCCs and LPCs, in figure 2, seems to saturate once the complexity of the background model reaches 2048 Gaussians. Such a behavior might be due to the small size of the speech corpus and we expect a higher saturation point in the case more training samples were available. At the same time, the poor performance depicted by pitch related features might be due to the few number of features employed. We extract pitch and log-pitch and the corresponding derivatives yielding to a 6-dimension feature vector whereas two other feature sets employ more than 40 dimensions.

Overall, results obtained in the experimental section sup-

port that modeling of dimensional emotions by means of binary fingerprint vectors obtains similar results to previous papers while it relies on a small set of frame-based acoustic features. We believe that the binary fingerprinting strategy, due its low-complexity and easy scalability, is a firm candidate to be used for a continuous and low cost approach in automatic emotion recognition systems, especially for low computational power devices.

4. CONCLUSIONS

A novel method has been presented in this paper for estimating natural vocal expressed emotions based on binary vectors modeling. Experiments reported on the VAM emotional speech corpus, in terms of regression coefficient and mean regression error, show that the binary fingerprint technique is able to successfully capture emotion traits that significantly correlate with ground-truth values across the three emotional dimensions analyzed. In this paper we have performed experiments with three different sets of input speech features. Results show that frame-based low-band MFCC acoustic features are good candidates to capture emotional cues in vocal expressed emotions. Furthermore, we have also reported preliminary results based upon the combination of binary models, which is an inexpensive and meaningful strategy to combine cues from different acoustic resolutions.

REFERENCES

[1] Shashidhar G. Koolagudi and K. Sreenivasa Rao, "Emotion recognition from speech: a review," *Int. J. Speech Technol.*, vol. 15, no. 2, pp. 99–117, 2012.

[2] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recogn.*, vol. 44, no. 3, pp. 572–587, 2011.

[3] Harold Schlosberg, "Three dimensions of emotion.," *Psychological Review*, vol. 61, no. 2, pp. 81–88, 1954.

[4] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Automatic Face Gesture Recognition and Workshops, IEEE International Conference on*, pp. 827–834, 2011.

[5] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10, pp. 787–800, 2007.

[6] M. Grimm, K. Kroschel, and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, vol. 4, pp. 1085–1088, 2007.

[7] F. Eyben, M. Wollmer, and B. Schuller, "Openear-2014; introducing the munich open-source emotion and affect recognition toolkit," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pp. 1–6, 2009.

[8] Florian Eyben, Anton Batliner, and Bjoern Schuller, "Towards a standard set of acoustic features for the processing of emotion in speech.," in *159th Meeting Acoustical Society of America,* vol. 9, 2012.

[9] E. Bozkurt, E. Erzin, C.E. Erdem, and A.T. Erdem, "Interspeech 2009 emotion recognition challenge evaluation," in *Signal Processing and Communications Applications Conference (SIU), IEEE 18th*, pp. 216–219 , 2010.

[10] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[11] Xavier Anguera and Jean-Franois Bonastre,"A novel speaker binary key derived from anchor models," in *INTERSPEECH, Proceedings on*, pp. 2118–2121, 2010.

[12] Esperanca Movellan, Xavier Anguera and Miquel Ferrarons, "Emotions recognition using binary fingerprints," in *Iberspeech, Proceedings on*, 2012.

[13] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[14] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *Multimedia and Expo, 2008 IEEE International Conference on*, pp. 865–868, 2008.

[15] Alex J. Smola and Bernhard Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[16] B. Milner and Xu Shao, "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 24–33, 2007.

[17] Paul Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *IFA Proceedings 17*, pp. 97–110, 1993.

[18] Daniel Neiberg, Kjell Elenius, and Kornel Laskowski, "Emotion recognition in spontaneous speech using gmms," in *Interspeech 9th International Conference on Spoken Language Processing*, pp. 809–812, ISCA-Inst Speech Communication Assoc, 2006.

[19] Ming Li, Kyu J. Han, and Shrikanth Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 151–167, 2013.

[20] Ming Li, Chi-Sang Jung, and Kyu Jeong Han, "Combining five acoustic level modeling methods for automatic speaker age and gender recognition.,' in *INTERSPEECH, Proceedings on*, pp. 2826–2829, 2010.