ITERATIVE FEATURE NORMALIZATION FOR EMOTIONAL SPEECH DETECTION

Carlos Busso

Multimodal Signal Processing (MSP)
Department of Electrical Engineering,
The University of Texas at Dallas
busso@utdallas.edu

Angeliki Metallinou and Shrikanth S. Narayanan

Speech Analysis and Interpretation Laboratory
Viterbi School of Engineering,
University of Southern California,
metallin@usc.edu, shri@sipi.usc.edu

ABSTRACT

Contending with signal variability due to source and channel effects is a critical problem in automatic emotion recognition. Any approach in mitigating these effects however has to be done so as to not compromise emotion-relevant information in the signal. A promising approach to this problem has been through feature normalization using features drawn from non-emotional ("neutral") speech samples. This paper considers a scheme for minimizing the inter-speaker differences while still preserving the emotional discrimination of the acoustic features. This can be achieved by estimating the normalization parameters using only neutral speech, and then applying the coefficients to the entire corpus (including emotional set). Specifically, this paper introduces a feature normalization scheme that implements these ideas by iteratively detecting neutral speech and normalizing the features. As the approximation error of the normalization parameters is reduced, the accuracy of the emotion detection system increases. The accuracy of the proposed iterative approach, evaluated across three databases, is only 2.5% lower than the one trained with optimal normalization parameters, and 9.7% higher than the one trained without any normalization scheme.

Index Terms— emotion recognition, fundamental frequency, emotions, feature normalization

1. INTRODUCTION

Advances in automatic emotion recognition systems can have a positive impact on the development of new human behavioral computing (behavioral informatics) capabilities as well as in the design of novel human machine interfaces. Some of the applications that could benefit from cognitive interfaces that are able to sense and react to the emotional state of the users include customer service, video games, education and health informatics. Unfortunately, the broad translation of the promising results shown on early stages of automatic emotion recognition to real life applications is still forthcoming mainly due to robustness of the methods to varying conditions and contexts [1]. One of the specific challenges in detecting the emotional state of the users is the ability to tackle the inter-speaker variability observed in expressive speech [2]. This is the focus of this paper.

Human speech production is the result of controlled anatomical movements of the lungs, trachea, larynx, pharyngeal cavity, oral cavity, and nasal cavity [3]. As a result, the properties of speech are intrinsically speaker dependent. Although there are patterns that are preserved across speakers (e.g., the increase of F0 in angry sentences), the expression of emotions presents idiosyncratic differences. In fact, previous works have shown that speaker dependent classifiers yield higher performance than speaker independent classifiers [4]. Therefore, it is not surprising that speech normalization can serve as a key step in building a robust emotion recognition system.

In this context, we propose a novel *iterative feature normalization* (IFN) scheme designed to reduce the speaker variability, while still preserving the signal information critical to discriminate between emotional states. The main idea of the normalization scheme considered here is to normalize the emotional corpus such that neutral speech from each speaker presents similar trends. For each speaker, the algorithm iteratively classifies the acoustic observations as either emotional or neutral. The speech samples that are recognized as neutral are used to estimate linear scaling parameters that are used for normalization. Then, these parameters are applied to the entire data, including the emotional subset, and the process is repeated. Since the normalization parameters are estimated using only the samples recognized as neutral, the discrimination between emotional and neutral classes is preserved. One notable advantage of using neutral speech samples for this purpose is the wide availability of neutral speech data.

The paper motivates and describes the iterative normalization algorithm as an extension of our previous work on detecting emotional speech using robust neutral reference models [5, 6]. The results show that the IFN scheme provides a good approximation of the optimal normalization parameters (estimated by assuming that emotional labels are known – perfect emotion detection). The accuracy achieved by the classifier using the IFN approach is only 2.5% lower than the one trained with optimal features normalization parameters, and 9.7% higher than the one trained without any normalization scheme. Although the IFN approach is presented in the context of detecting emotional speech with neutral models in this paper, the algorithm can be generalized to other multi-class emotion recognition problems.

2. MOTIVATION

2.1. Neutral model approach for emotion detection

We have previously proposed the use of neutral models to detect emotional speech (see Fig. 1) [5, 6]. The main idea is to train robust acoustic models such as hidden Markov models (HMMs) and Gaussian mixture models (GMMs) with neutral speech (say using a generic reference spontaneous speech corpus, different from the emotional databases). The observations are contrasted with these models and their likelihoods are used as *fitness measures*. For neutral speech, the models will fit the observation and the likelihood will be higher. Under the assumption that emotional speech differs from its neutral counterpart, it is expected that a lower likelihood will be observed for emotional speech. We have shown that classifiers trained with these likelihoods as features are more robust and

Table 1. Summary of the databases (*neu*=neutral, *ang*=anger, *hap*=happiness, *sad*=sadness, *bor*=boredom, *dis*=disgust, *fea*=fear, *anx*=anxiety, *pan*=panic, *anh*=hot anger, *anc*=cold anger, *des*=despair, *ela*=elation, *int*=interest, *sha*=shame, *pri*=pride, *con*=contempt

Data	type	Use of the data	Spontaneous/acted	Language	# speakers	# utterances	Emotions
WSJ1	neutral	Reference	spontaneous	English	50	8104	neu
EMA	emotion	Training/testing	acted	English	3	688	neu,ang,hap,sad
EMO-DB	emotion	Training/testing	acted	German	10	535	neu,ang,hap,sad,bor,dis,fea
EPSAT	emotion	Training/testing	acted	English	8	4738	neu,hap,sad,bor,dis,anx,pan,anh,anc,des,ela,int,sha,pri,con

generalize better than the direct use of the acoustic features. This approach can even detect expressive speech from emotional classes that were not included in the training set, as long as its acoustic features differ in any aspect from neutral speech.

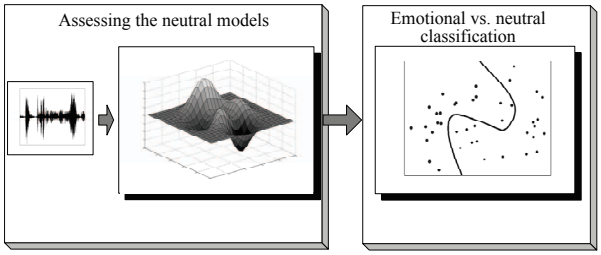


Fig. 1. A two-step approach to discriminate neutral versus emotional speech. In the first step, the input speech is contrasted with robust neutral references models. In the second step, the *fitness measures* are used for binary emotional classification.

2.2. Optimal feature normalization

One important step in this approach is to normalize acoustic features. The task is to reduce speaker variability, while preserving the discrimination between emotional classes. This goal is achieved by reducing the differences observed in the neutral subsets across speakers. The normalization parameters, which are estimated from the neutral subset, are applied to the entire emotional corpus, including the emotional subset. Therefore, the differences between emotional classes are preserved. This procedure is repeated for each speaker. This approach has been used to normalize not only acoustic features, but also facial motion units [7].

An implementation of this approach was presented in our previous work to normalize the F0 mean [6]. The average pitch across speakers in the neutral reference database was estimated, $F0_{ref}$ (reference corpus used to train the neutral models). Then, the average pitch value for the neutral set of the emotional databases was estimated for each speaker, $F0^s_{neu}$. Finally a scaling factor (S^s_{F0}) was calculated by taking the ratio between $F0_{ref}$ and $F0^s_{neu}$, as shown in Equation 1. Therefore, the neutral samples of each speaker in the databases will have a similar F0 mean value.

$$S^s_{F0} = \frac{F0_{ref}}{F0^s_{neu}} \quad (1)$$

This normalization scheme makes two important assumptions: (1) A portion of neutral speech from each speaker is available to estimate the normalization parameters (S^s_{F0}). (2) The identities of the speakers in the emotional databases are known.

These two assumptions restrict the use of this approach in practical applications, in which speech with emotional labels from users is not available during training. In this context, the IFN approach is designed to address and relax the first assumption by iteratively normalizing the data and detecting neutral speech. It is still assumed that the identities of the speakers are known. There are various applications in which it is reasonable to assume that speech from a single speaker is recorded (e.g., call centers). In these cases, the proposed IFN approach provides an ideal normalization scheme.

3. ITERATIVE FEATURE NORMALIZATION

The purpose of the *iterative feature normalization* (IFN) approach is to successively detect neutral speech from the database and estimate the normalization parameters using only this subset. The goal of the approach is two fold: to reduce the differences between the neutral subsets across speakers, and preserve the discrimination between emotional and neutral speech.

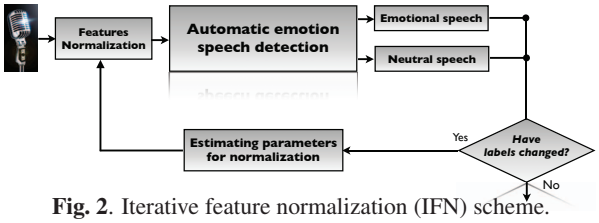


Fig. 2. Iterative feature normalization (IFN) scheme.

Figure 2 describes the algorithm. First, acoustic features without any normalization are used to detect expressive speech (neutral versus emotional classes). The observations that are labeled as neutral are used to re-estimate the normalization parameters. As the approximation of the normalization parameters improves, the performance of the detection algorithm is expected to improve, leading to better normalization parameters. The process is repeated until the percentage of files in the emotional database that change labels from successive iterations is lower than a given threshold (e.g., 5%). As discussed in section 5.1, the convergence state of the algorithm is conditioned by the performance of the emotion detection algorithm (Fig. 2). In this paper, this emotion detector is implemented with the neutral model approach described in section 2.1.

Although the IFN framework is implemented in the context of detecting emotional speech using neutral models, the ideas behind the approach can be generalized and extended to other emotion recognition problems.

4. IMPLEMENTATION & EXPERIMENTAL SETUP

4.1. IFN implementation

The IFN approach is implemented in the context of our previous work on emotional speech detection using prosodic features [6]. In that work, the discriminative information of statistics derived from the fundamental frequency contour was analyzed. Seven features were selected to train the neutral models described in section 2.1: the upper (*SQ75*) and lower quartiles (*SQ25*), median of the F0 contour (*Smedian*) and its derivative (*Sdmedian*), interquartile range of F0 derivative (*Sdiqr*), mean of the voiced segment range (*SVmeanRange*), and maximum value of the voiced segment curvatures (*SVmaxCurv*). Details about these features are provided in [6]. These are also the features used in the experiments reported here.

Although the approach can be implemented to normalize any aspect of acoustic features, this paper focuses on F0 mean. The fundamental frequency is directly constrained by the structure and size of the larynx [3]. Therefore, the F0 mean presents strong inter-speaker variability [2], which makes it a good candidate to validate the proposed IFN approach. After the fundamental frequency is estimated, it is scaled by the factor S^s_{F0}, as described in section 2.1. The aforementioned statistics are then derived from the scaled F0 contour.

A threshold on the classifier likelihood is used to improve the precision of the neutral subset used to estimate the normalization coefficients. Only observations with likelihood of being neutral higher than a given threshold are used for this purpose. Preliminary analysis suggested that setting this threshold equal to 0.7 yields to enough neutral samples, maintaining an acceptable accuracy rate. This subset was expanded to include enough data for each subject. When the percentage of selected speech files for one subject was lower than 20% (empirically chosen) of his/her samples in the corpus, files with the highest likelihood were added to the set, even when the aforementioned threshold was not satisfied.

4.2. Databases and emotion detection classifiers

Similar to the experimental set up used in our previous work [6], three emotional databases are used: USC-EMA [8], EMO-DB [9], and EPSAT [10]. Table 1 gives details about these corpora. The databases are jointly used to train the classifiers. The samples for the various specific emotional classes are considered broadly as emotional.

For the neutral models (Sec. 2.1, Fig. 1), a GMM for each F0 statistic was trained using the Wall Street Journal-based Continuous Speech Recognition Corpus Phase II (WSJ) [11] (see Table 1). The likelihoods of the models are used as features in the classification step. For classification, a Linear Discriminant Classifier (LDC) was implemented to detect emotional from neutral speech.

Since the emotional categories are grouped together, the number of emotional samples is higher than the neutral samples in the three databases above. Therefore, the emotional samples were randomly drawn to match the number of neutral samples (baseline 50%). This process was repeated 400 times. The recognition results presented here correspond to average values over these 400 realizations. Then, the selected samples were split in training and testing sets (70% and 30%, respectively). Notice that the three emotional corpora are considered together. For comparison, a classifier was implemented with LDC using directly the F0 statistics, instead of the likelihood of the GMM neutral models, which is referred to as the *conventional scheme*.

5. RESULTS

Table 2 gives the performance of the emotional speech classifiers using the neutral model and conventional approaches when different normalization schemes are used. For each condition, the overall performance is provided (*All*). In addition, the results are disaggregated in terms of the emotional databases. Notice that the 3 emotional databases are jointly used for training and testing.

Table 2 shows the performance of the classifiers when the F0 contour is scaled with the actual normalization coefficients (*optimal normalization*). It also shows the performance when no normalization scheme is used (*without normalization*). The results show that the accuracy decreases in 12.2% for the neutral model approach, and 6.9% for the conventional approach. This result suggest that feature normalization can serve as a key step in any emotion recognition system.

Table 2. Performances of the neutral model, and conventional approaches with different normalization schemes (*Acc*=Accuracy, *Rec*=Recall, *Pre*=Precision, *F*= F - score).

	Neutral Model				Conventional scheme			
Optimal Normalization [%]								
	Acc	Rec	Pre	F	Acc	Rec	Pre	F
All	**78.1**	**80.2**	**74.6**	**77.3**	**74.6**	**89.5**	**55.7**	**68.7**
EMA	86.6	92.1	71.9	80.8	81.7	95.2	56.0	70.5
EMO-DB	80.5	85.9	77.4	81.4	77.6	94.6	63.0	75.6
EPSAT	74.9	76.7	74.6	75.7	71.7	87.1	54.0	66.7
Without normalization [%]								
	Acc	Rec	Pre	F	Acc	Rec	Pre	F
All	**65.9**	**65.1**	**69.0**	**67.0**	**67.7**	**72.8**	**56.6**	**63.7**
EMA	73.9	72.7	54.2	62.1	68.7	64.5	45.8	53.6
EMO-DB	66.1	69.7	68.3	69.0	72.4	84.8	61.0	71.0
EPSAT	63.4	63.1	72.8	67.6	66.4	72.2	58.4	64.6
Global normalization (speaker dependent) [%]								
	Acc	Rec	Pre	F	Acc	Rec	Pre	F
All	**65.8**	**66.2**	**64.4**	**65.3**	**72.1**	**82.2**	**56.4**	**66.9**
EMA	73.6	69.1	57.5	62.8	77.2	72.4	66.7	69.4
EMO-DB	70.7	81.8	59.8	69.1	75.4	89.6	62.5	73.6
EPSAT	62.3	63.4	67.0	65.2	69.8	83.9	52.7	64.7
IFN approach (speaker dependent) [%]								
	Acc	Rec	Pre	F	Acc	Rec	Pre	F
All	**75.6**	**76.6**	**73.7**	**75.1**	**73.3**	**86.7**	**55.0**	**67.3**
EMA	85.2	88.5	71.1	78.8	80.8	97.8	51.9	67.8
EMO-DB	74.1	77.9	73.9	75.8	76.8	92.7	62.8	74.8
EPSAT	72.8	74.1	74.2	74.2	70.2	83.3	54.1	65.6

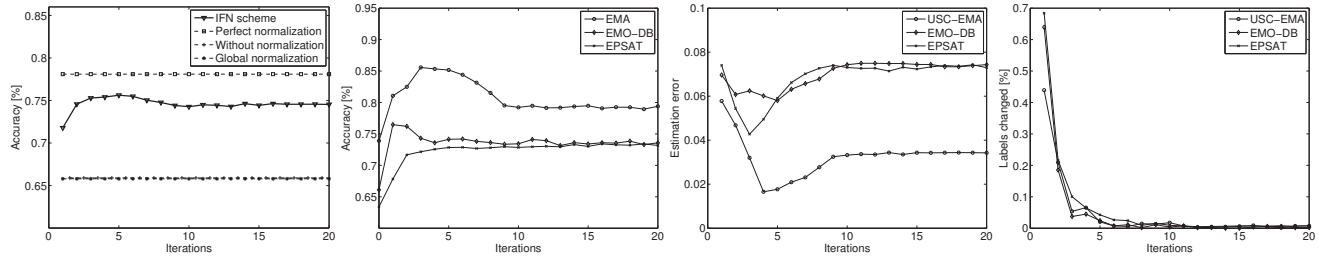
Table 2 also shows the performance when normalization parameters are estimated for each speaker using all the samples, including neutral and emotional subsets (*global normalization*). Since the parameters are estimated for each speaker, the knowledge of the identity of the files is still required (speaker dependent normalization). The accuracy of the system is not improved by this normalization scheme, specially for the neutral model approach. This result indicates that global normalization, which is the most common normalization approach, affects the discriminative power of the acoustic features.

Table 2 and Figure 3 give the results for the proposed IFN approach. At the fifth iteration, the classifier reached the best accuracy, which is only 2.5% lower and 9.8% higher than accuracies achieved by using optimal normalization and global normalization, respectively. Notice that the IFN approach also improves the accuracy of the conventional classifier, which is only 1.3% lower than that achieved with optimal feature normalization. These results show the potential of the proposed IFN approach.

The improvement in the performance for the EMO-DB database is not as impressive as the improvement in other databases. Figure 3-b even shows that the accuracy decreases in early iteration, until it finally converges at 73.6%. This is the database with the lowest average number of samples per speaker (53.5), which may explain this result.

5.1. Convergence and stopping criteria

Figure 3-d shows the percentage of labels assigned to the emotional data that are changed from consecutive iterations. As can be seen, after the fifth iteration less than 5% of the labels changed. Furthermore, Figure 3-c indicates that the estimation error, calculated as the average of the absolute difference between the optimal and estimated

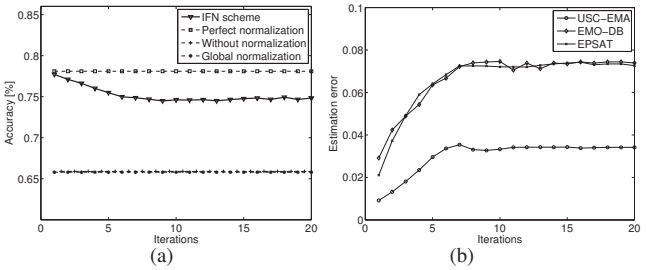


(a) (b) (c) (d)

Fig. 3. Performance of the IFN approach. (a) Accuracy of the emotion detection system, (b) Accuracy of the emotion detection system for each database – iteration 0 gives the accuracy without IFN, (c) Approximation of normalization coefficients, (d) Percentage of files that change emotional labels from previous iteration.

parameters, also converges to a stable value. After some iterations, the detected neutral subset is only slightly modified. These empirically illustrate the convergence potential of the IFN algorithm.

Given that the performance of the expressive speech detection system is not perfect, the IFN approach will only approximate the real values of the normalization coefficients. As a result, the performance of the IFN approach may not reach the upper bound imposed by training the classifier with optimal normalization coefficients.



(a) (b)

Fig. 4. Performance of the system when the normalization parameters are initialized with the true values. (a) Accuracy of the emotion detection system (b) Approximation of normalization coefficients.

To analyze the convergence of the IFN approach, the normalization parameters were initialized using optimal normalization parameters. The results are given in Figure 4. Figure 4-b shows that the error estimation in the normalization parameters increases because of the misclassification produced by the system. As a result, the algorithm will converge to a suboptimal state yielding lower accuracy (from 78.1% to 74.5%). This result is also observed in Figures 3-b and 3-c. Notice that this accuracy is still 8.7% higher than that achieved with global normalization.

6. CONCLUSIONS

The paper introduced an *iterative feature normalization* scheme for emotion detection. By iteratively detecting neutral speech and normalizing the acoustic features, the IFN scheme approximates the optimal normalization parameters, which minimizes the differences between the neutral subset of each speaker while still preserving the emotional discrimination conveyed in the speech data. The accuracy of the emotion detection system trained with the IFN approach is only 2.5% lower than the one trained with optimal features normalization parameters, and 9.8% higher than the one trained with global normalization parameters.

One limitation of the IFN approach is that it assumes that the identities of the speakers are known. Depending on the problem, this assumption can be relaxed by estimating the identity using supervised or unsupervised speaker identification. Another direction that we are exploring is the use of this approach in other multiclass emotion recognition tasks. It is expected that this approach will prove to be useful in improving the performance and robustness of broad classes of emotion classifiers.

The results presented in this paper suggest that the IFN approach has the potential to reduce inter-speaker variability. This will move us closer toward cognitive interfaces able to robustly sense the emotional states of the users.

7. REFERENCES

[1] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "Desperately seeking emotions or: actors, wizards and human beings," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, Northern Ireland, UK, September 2000, pp. 195–200.

[2] C. Busso, M. Bulut, and S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, S. M. J. Gratch, Ed. New York, NY, USA: Oxford University Press, 2010.

[3] T. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ: Prentice Hall, November 2001.

[4] B. Schuller, R. Müller, M. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, Lisbon, Portugal, September 2005, pp. 805–808.

[5] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," in *Interspeech 2007 - Eurospeech*, Antwerp, Belgium, August 2007, pp. 2225–2228.

[6] ——, "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, May 2009.

[7] Z. Zeng, J. Tu, B. Pianfetti, and T. Huang, "Audiovisual affective expression recognition through multistream fused HMM," *IEEE Transactions on Multimedia*, vol. 10, no. 4, pp. 570–577, June 2008.

[8] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, "An articulatory study of emotional speech production," in *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, Lisbon, Portugal, September 2005, pp. 497–500.

[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, Lisbon, Portugal, September 2005, pp. 1517–1520.

[10] M. Liberman, K. Davis, M. Grossman, N. Martey, and J. Bell, "Emotional prosody speech and transcripts," Philadelphia, PA, USA, 2002, Linguistic Data Consortium.

[11] D. Paul and J. Baker, "The design for the Wall Street Journal-based CSR corpus," in *2th International Conference on Spoken Language Processing (ICSLP 1992)*, Banff, Alberta, Canada, October 1992, pp. 899–902.