Confidence Measures in Speech Emotion Recognition Based on Semi-supervised Learning

Jun Deng[1], *Björn Schuller*[1]

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
{jun.deng, schuller}@tum.de

Abstract

Even though the accuracy of predictions made by speech emotion recognition (SER) systems is increasing in precision, little is known about the confidence of the predictions. To shed some light on this, we propose a confidence measure for SER systems based on semi-supervised learning. During the semi-supervised learning procedure, five frequently used databases with manually created confidence labels are implemented to train classifiers. When the SER system predicts the label for an unknown test utterance, these classifiers serve as a reliability estimator for the utterance and output a series of confidence ratios that are combined into a single confidence measure. Our experimental results impressively show that the proposed confidence measure is effective in indicating how much we can trust the predicted emotion.

Index Terms: speech emotion recognition, confidence measure, semi-supervised learning, cross-corpus

1. Introduction

When speech emotion recognition (SER) systems are employed in any real-world application, it is common that many decisions made by the systems are not always correct, even for the best SER systems available today. This may be due to ambient noise, database divergence (e. g., acted/non-acted, induced, and naturalistic databases), redundant and correlated features, and many other negative influence factors from the real world. The ability to assess reliability or probability of correctness of every decision is thus essential to increase usefulness and "intelligence" of a SER system.

A number of approaches for the topic of Confidence Measures (CM) have been proposed in the domain of Automatic Speech Recognition (ASR). Generally speaking, these approaches can be roughly grouped into three categories [1]: in the first, a two-class (*true* or *false*) classifier is built based on a combination of the so-called predictor features (e. g., acoustic stability and language model scores) collected during the decoding procedure. Various classification models have been used in the literature for this purpose, including the linear discriminant function[2], maximum entropy model [3], etc. In the second category, an approximation of the posterior probability in the standard maximum a posteriori (MAP) criterion approach is taken as the confidence measure. The posterior probability is typically estimated from the speech system lattices or N-best lists [4]. Methods in the third category treat the confidence estimation problem as an utterance verification problem. They make use of the likelihood ratio between the null hypothesis (e. g., the word is correct) and the alternative hypothesis (e. g., the word is incorrect) as a confidence measure [5]. Unfortunately, the use of CM for SER systems seems to have never drawn any attention so far. All the existing approaches are primarily designed for ASR systems and not for SER systems, as most of them rely almost entirely on properties of the Hidden Markov Models (HMM), such as acoustic scores or the word graph, which are not typical components of SER systems.

In this study, we aim to make SER systems have this "intelligence" by proposing CM based on semi-supervised learning. The semi-supervised learning algorithm is implemented to iteratively train the classifiers from other corpora which gradually include data from the training set of the original corpora. The agreement of these classifiers are combined to calculate the CM for every instance. For the task of emotion recognition, we concentrate on the INTERSPEECH 2009 Emotion Challenge two-class task using the FAU Aibo emotion corpus for evaluation.

In the remainder of this paper, we briefly describe the six selected emotional speech databases (Section 2). We further show the details of our proposed CM using semi-supervised learning in Section 3. Then, we present the results in Section 4. Finally, we conclude our study in Section 5.

2. Databases

For our study we chose six among the most frequently used databases in the field. The content ranges from acted over induced to spontaneous affect portrayal. For better comparability among corpora, we map the diverse emotion classes onto one of the two most popular axes in the dimensional emotion models: valence (i. e., negative ("-") vs. positive ("+")). In the following, each database is briefly introduced along with the mapping of classes to binary valence ("+" and "-") and the number of instances.

The Airplane Behaviour Corpus (ABC) is crafted for the application of public transport surveillance. It is based on induced mood by pre-recorded announcements of a vacation (return) flight, consisting of 13 and 10 scenes. And it contains aggressive (-, 95), cheerful (+, 105), intoxicated (-, 33), nervous (-, 93), neutral (+, 79), and tired (-, 25) speech.

The Audiovisual Interest Corpus (AVIC) is made up of spontaneous speech and natural emotion. In its scenario setup, a product presenter leads subjects through an English commercial presentation. AVIC is annotated in "level of interest" (loi) from low (1) to high (3). Loi2 (279 instances) and loi3 (170 instances) are mapped to positive and loi1 (553 instances) to negative valence.

The Danish Emotional Speech (DES) database contains nine professionally acted Danish sentences, two words, and chunks that are located between two silent segments of two passages of fluent text. Emotions include angry (-, 85), happy (+, 86), neutral (+, 85), sadness (-, 84), and surprise (+, 79).

The eNTERFACE database (eNTER) consists of recordings

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

Corpus	Language	Speech	Emotion	# Valence		# All	h:mm	# m	# f	# Lab	Rec	kHz
				-	+							
ABC	German	fixed	acted	213	217	430	1:15	4	4	3	studio	16
AVIC	English	free	natural	553	2 449	3 002	1:47	11	10	4	studio	44
DES	Danish	fixed	acted	169	250	419	0:28	2	2	–	studio	20
eNTER	English	fixed	induced	855	422	1 277	1:00	34	8	2	studio	16
SAL	English	free	natural	917	779	1 692	1:41	2	2	4	studio	16
FAU Aibo	German	free	natural	5 823	12 393	18 216	9:20	21	30	5	noisy	16

of native subjects from 14 nations speaking pre-defined spoken content in English. Each subject listened to six successive short stories designed to elicit a particular emotion out of anger (-, 215), disgust (-, 215), fear (-, 215), happiness (+, 207), sadness (-, 210), and surprise (+, 215).

The Belfast Sensitive Artificial Listener (SAL) data is part of the HUMAINE database. The set contains recordings of natural human-SAL conversations from 4 speakers, with an average length of 20 minutes per speaker. The data has been labeled continuously in real time with respect to valence and activation using a system based on FEELtrace. The annotations were globally normalized to zero mean and scaled so that 98 % of all values are in the range from -1 to +1. The 25 recording sessions have been split into chunks using an energy based Voice Activity Detection. Labels for each chunk are computed as average of the continuous labels within the chunk. There are 779 chunks with positive and 913 chunks with negative valence.

The FAU Aibo emotion corpus contains emotionally coloured, spontaneous, German speech. We use the labels of the INTERSPEECH 2009 Emotion Challenge two-class task and map **NEG** to negative valence, and **IDL** to positive valence. Thus, in the training set there are 6 601 instances of positive and 3 358 instances of negative valence, and in the test set we have 5 792 instances of positive valence and 2 465 instances of negative valence.

Details on the corpora are summarized in Table 1 and more information is found in [6]. Note that in the ongoing, balancing of the training partition is used by random upsampling to unity.

Table 2: 39 functionals used in the TUM openEAR emo_large set.

Functionals	#
Respective rel. position of max./min. value	2
Range (max.-min.)	1
Max. and min. value - arithmetic mean	2
Arithmetic mean, Quadratic mean, Centroid	3
Number of non-zero values	1
Geometric, and quadratic mean of non-zero values	2
Mean of absolute values, Mean of non-zero abs. values	2
Quartiles and inter-quartile ranges	6
95 % and 98 % percentile	2
Std. deviation, variance, kurtosis, skewness	4
Zero-crossing rate	1
# of peaks, mean dist. btwn. peaks, arth. mean of peaks, arth. mean of peaks - overall arth. mean	4
Linear regression coefficients and error	4
Quadratic regression coefficients and error	5

Table 3: 56 Low-Level Descriptors used in the TUM openEAR emo_large set.

Feature Group	Features in Group
Raw Signal	Zero-crossing-rate
Signal energy	Logarithmic
Pitch	Fundamental frequency F_0 in Hz via Cepstrum and Autocorrelation (ACF). Exponentially smoothed F_0 envelope.
Voice Quality	Probability of voicing ($\frac{ACF(T_0)}{ACF(0)}$)
Spectral	Energy in bands 0–250 Hz, 0–650 Hz, 250–650 Hz, 1–4 kHz 25 %, 50 %, 75 %, 90 % roll-off point, centroid, flux, and rel. pos. max. / min.
Mel-spectrum	Band 1–26
Cepstral	MFCC 0–12

2.1. Acoustic Features

Over the last few years, the focus has been placed increasingly on acoustic features. We decided on a typical state-of-the art emotion recognition engine operating on supra-segmental level, and use our open source openEAR toolkit [7] to extract a set of systematically generated acoustic features. We employ the pre-defined openEAR configuration "emo_large" with 39 functionals of 56 acoustic Low-Level Descriptors (LLDs) and their first and second order delta regression coefficients. The 39 statistical functionals and 56 LLDs used are shown in Table 2 and 3 respectively. This results in a total of 6 552 acoustic features.

3. CM using semi-supervised learning

A straightforward method to understand the reliability of a prediction made by a classification system is to consider its confusion matrix. A confusion matrix includes information about actual and predicted classifications made by a classification system. For a two-class problem of positive and negative valence, Table 4 shows a generic confusion matrix. If the prediction of an instance made by this classification system is negative, for example, the ratios $a/(a+c)$ and $c/(a+c)$ can roughly indicate the reliability of the instance correctly classified and incorrectly classified as negative based on its confusion matrix, respectively. There are four type of ratios for the two-class problem. We call them confidence ratios defined as follows:

$$\begin{aligned} CN = \tfrac{a}{a+c} \quad ICN = \tfrac{c}{a+c} \\ CP = \tfrac{d}{b+d} \quad ICP = \tfrac{b}{b+d} \end{aligned} \tag{1}$$

where the *CN* (Correct Negative) ratio is the proportion of negative predictions correctly classified; the *ICN* (Incorrect Nega-

Table 4: Generic confusion matrix for the 2-class problem.

	Predicted -	**Predicted +**
Actual -	a	b
Actual +	c	d

tive) ratio is the proportion of positive predictions incorrectly classified as negative; the *CP* (Correct Positive) ratio is the proportion of positive predictions correctly classified; *ICP* (Incorrect Positive) ratio is the proportion of negative predictions incorrectly classified as positive.

In general, a combination approach to describe the reliability of a prediction by individual components can usually improve the overall performance. To obtain various sets of the confidence ratios, therefore, the fixed number of corpora are used as test sets to deal with the recognition task on the original training set for emotion recognition, respectively. Then, a new class label (one of CN, ICN, CP, and ICP) is assigned for each instance in the selected databases.

Table 5: Training procedure of the proposed semi-supervised learning algorithm.

```
Inputs:
  $N$ is the number of the selected corpora;
  $N$ sets of instances: $L_i, (i = 1, 2, ..., N)$;
  $N$ sets of predictions: $P_i, (i = 1, 2, ..., N)$;
  A set of high confidence instances: $T = \emptyset$;
  A set of unlabeled instances: $U$;
Process:
  Build $N$ classifiers: $h_i(L_i), (i = 1, 2, ..., N)$;
  While iteration < MAXITERATION
    Predictions: $P_i = h_i(U), (i = 1, 2, ..., N)$;
    Pick high confidence instances from $U$ according to
    majority voting $P_i$:
    $T$ = MV($P_1, ..., P_N$);
    If $T$ is empty
      Break;
    End
    $L_i = L_i \cup T, (i = 1, 2, ..., N)$;
    $U_i = U_i - T, (i = 1, 2, ..., N)$;
    Rebuild the classifiers: $h_i(L_i), (i = 1, 2, ..., N)$;
  End
Outputs:
  $N$ trained classifiers: $h_i, (i = 1, 2, ..., N)$.
```

When the selected corpora together with the newly assigned labels are employed to train classifiers which estimate the reliability of each recognition decision made by a SER system, we develop a semi-supervised learning algorithm to iteratively train these classifiers and reinforce their prediction correction. In this method, we consider the original training set for emotion recognition to be a set of unlabeled instances when learning. The semi-supervised learning algorithm proceeds as shown in Table 5. Given N is the number of the selected corpora, let $L_i(i = 1, 2, ..., N)$ be the selected corpora i with the newly assigned labels (i. e, CN, ICN, CP, and ICP), P_i be a set of predictions generated by a classifier $h_i(L_i)$ trained on database $i(i = 1, 2, ..., N)$, and U be the unlabeled set of the training set for emotion recognition task. At each iteration, the unlabeled set U is predicted by classifier $h_i(L_i)(i = 1, 2, ..., N)$, respectively. A set of high confidence instances T is a subset of U containing only instances with high confidence predictions. In order to obtain the set T, the majority voting (MV) rule is carried out: If the predictions of three or more classifiers $h_i(L_i)$ agree, the instance with the majority vote label assigned is included in T. Afterward, the high confidence set T is joined with the training set $L_i(i = 1, 2, ..., N)$ and is removed from the unlabeled set U at the same time. The classifiers h_i are re-trained and the procedure is repeated until the confidence set T is empty or the number of iterations exceeds the maximum desired number. After the final iteration, we investigate the number of four newly assigned labels in each training set L_i and calculate confidence ratios in accordance with Equation (1).

While recognizing an instance from the FAU Aibo test set by the classifier trained on the FAU Aibo training set and by classifiers h_i at final iteration, all the predictions are used to estimate a final CM for the instance. Note that there are four types of predictions from the classifiers h_i in total, they will directly represent four types of confidence ratios when calculating a final CM. To guarantee the final CM to be between 0 and 1, a normalization is implemented on the confidence ratios. Given a predicted emotion hypothesis e for an test instance and the corresponding confidence ratios $r_i \in \{CN, ICN, CP, ICP\}(i = 1, 2, ..., N)$, the normalized confidence ratios can be defined as follows:

$$\tilde{r}_i = \begin{cases} \frac{r_i + \sum_{i=1}^{N} max(r_i^{CP}, r_i^{ICP})}{A} & \text{if } r_i \text{ is } CN \\ & \quad \text{or } ICN, \\ \frac{r_i + \sum_{i=1}^{N} max(r_i^{CN}, r_i^{ICN})}{A} & \text{otherwise} \end{cases} \tag{2}$$

where

$$A = \sum_{i=1}^{N} (max(r_i^{CN}, r_i^{ICN}) + max(r_i^{CP}, r_i^{ICP})). \tag{3}$$

Finally, a CM S_{CM} is calculated by:

$$S_{\text{CM}} = \sum_{i=1}^{N} w_i \times \tilde{r}_i \tag{4}$$

where w_i is a binary weighted factor; $w_i = +1$ if r_i has strong correlation with e (i. e., CN or ICN with negative valence) and $w_i = -1$ otherwise (i. e., CP or ICP with negative valence); $S_{\text{CM}} \in [0, 1]$. For a prediction made by this SER system with the S_{CM}, it is easy to know how likely it is correctly recognized. The higher S_{CM} of a prediction is , the more confident one is.

4. Evaluation

We evaluate our approach on the FAU Aibo emotion corpus, and use the corpora ABC, AVIC, DES, eNTER, and SAL, in the described semi-supervised learning approach to train classifiers for estimating CMs. As classifier, we consider Support Vector Machines (SVM) with polynomial kernel using sequential minimal optimization for training with complexity 0.01. For the two-class task with SVM and the emo_large feature set, the unweighted and weighted accuracies of the SER system are 68.7 % and 69.2 %, respectively. Table 6 comprises the confidence ratios of the five trained classifiers obtained by the proposed semi-supervised learning algorithm.

Normally, once the CM has been computed, each recognized emotion is simply annotated as either *correct* or *false*. When evaluating CM annotation, we usually encounter two

Table 6: Confidence ratios of the five trained classifiers obtained by the proposed semi-supervised learning algorithm.

Training set	**ABC**		**AVIC**		**DES**		**eNTER**		**SAL**	
Predicted	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**
Actual -	0.7926	0.1140	0.5644	0.0694	0.6997	0.1083	0.7130	0.2050	0.7123	0.2192
Actual +	0.2074	0.8860	0.4356	0.9306	0.3003	0.8917	0.2870	0.7950	0.2877	0.7808

types of errors, namely false alarm errors and false rejection errors. Obviously, the receiver operating characteristic (ROC) curve gives a full picture of performance at all operating points [1]. As a reference for CM, we consider the random guess method that every decision is randomly assigned a CM between 0 and 1. Figure 1 shows the comparison of the proposed semi-supervised learning method with different iteration stages and random guess in the ROC space. From Figure 1, it is easy to see that the proposed method produces a quick convergence in three iterations. Moreover, the ROC obtained by this method for CMs is significantly higher than by random guess, and its area under curve (AUC) benefits from the process of iterated semi-supervised learning and gains the maximum of $66.43\,\%$ at iteration 3 which is much higher than the AUC of random guess at $50\,\%$.

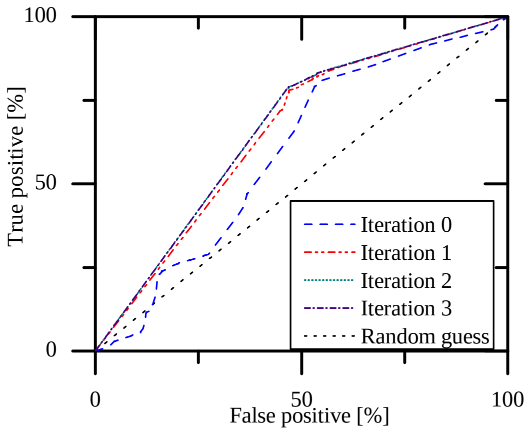

Figure 1: Comparison of the proposed semi-supervised learning method with different iteration stages and random guess in the ROC space. Random guess: Every decision is randomly assigned a CM between 0 and 1.

In many cases, it is convenient to use a single-number metric for CM assessment. Normalized Cross Entropy (NCE) is widely used [1], which is defined as

$$\mathrm{NCE} = \frac{H_{\text{base}} - H_{\text{cond}}}{H_{\text{base}}} \tag{5}$$

where

$$H_{\text{cond}} = -\sum_{i=1}^{N} \log(S_{\text{CM}}^{i}\delta(z_i = 1) + (1 - S_{\text{CM}}^{i})\delta(z_i = 0)) \tag{6}$$

and

$$H_{\text{base}} = -n\log(\frac{n}{N}) - (N-n)\log(1 - \frac{n}{N}). \tag{7}$$

Here, we know a set of N_t confidence scores and the associated class labels $\{(S_i \in [0,\,1], z_i\{0,\,1\})|i = 1, ..., N_t\}$, where N_t is the number of test instances, and $z_i = 1$ if the recognition result is correct and $z_i = 0$ otherwise. In Equation (6), $\delta(x) = 1$ if x is true and $\delta(x) = 0$ otherwise, and n is the number of samples whose $z_i = 1$.

The higher the NCE is, the better the CM quality. An optimal NCE equals one when a CM always outputs one when the recognition result is correct, and zero otherwise. Compared to the NCE of random guess -1.32, the proposed method for the FAU Aibo reaches a value of 0.80 which is close to the optimal value of 1, showing that the proposed method is effective in indicating the reliability of decisions from the SER system.

5. Conclusions

In this paper, we proposed a novel CM for a SER system based on semi-supervised learning. A self-training procedure was presented to utilize a given set of corpora to train classifiers for automatic confidence assignment. These classifiers led to a reliable way of assessing the correctness of the decisions of the SER system. The experimental results evaluated on the INTERSPEECH 2009 Emotion Challenge two-class problem demonstrated that our method obtained better performance compared to random guess probability as confidence measure. Moreover, it achieved a very good Normalized Cross Entropy value of 0.80, demonstrating its effectiveness. In the future, we would like to use the confidence measures to conduct semi-supervised learning on unlabeled emotional speech data.

6. Acknowledgements

Jun Deng's work is supported by the China Scholarship Council. The authors wish to thank Zixing Zhang and Florian Eyben for their suggestions.

7. References

[1] H. Jiang, "Confidence measures for speech recognition: a survey," *Speech Communication*, vol. 45, no. 4, pp. 455–470, 2005.

[2] R. Sukkar, "Rejection for connected digit recognition based on gpd segmental discrimination," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. 1. IEEE, 1994, pp. I–393–I–396.

[3] C. White, J. Droppo, A. Acero, and J. Odell, "Maximum entropy confidence estimation for speech recognition," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. 809–812.

[4] F. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 3, pp. 288–298, 2001.

[5] M. Rahim, C. Lee, and B. Juang, "Discriminative utterance verification for connected digits recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 5, no. 3, pp. 266–277, 1997.

[6] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 552–557.

[7] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit," in *Proc. ACII*, Amsterdam, 2009, pp. 576–581.