Confidence Measures in Multiclass Speech Emotion Recognition using Ensemble Learning to Catch Blunders

Alan Murphy
PhD Candidate
Department of Information Technology
National University of Ireland Galway

Sam Redfern
Lecturer
Department of Information Technology
National University of Ireland Galway

Abstract

Although Speech Emotion Recognition (SER) systems have continually been improved with regard to outputting decisions on class membership, it is only very recently that Confidence Measures (CM) have been incorporated into such systems. It is easy to presume that classifiers such as k-Nearest Neighbor, Naïve Bayes or Support Vector Machines that can readily output numeric distributions, can therefore easily produce confidence estimates. However these numeric outputs have proven not to be well correlated with classification confidence. With this in mind the contribution of this paper is threefold, 1) providing the first successful demonstration of confidence extraction in the multiclass problem of SER by using metrics from a k-NN meta-classifier within an ensemble architecture . 2) We also demonstrate how this system has shown improvement on a similar approaches discussed later, and 3) finally we note how confidence measures can be useful as a tool for catching blunderous predictions made in SER problems, which can be very useful in real world deployed systems.

Keywords: Confidence Measures, Emotion Recognition, Ensemble Learning, Multiclass Speech Classification

I. INTRODUCTION

Previous research into confidence in speech classification predictions has been predominantly designed for Automatic Speech Recognition (ASR) applications and not for SER systems. Only one paper in the speech processing community has previously dealt with the latter [1]. These research endeavors utilize components such as acoustic scores and acyclic word lattices or graphs which are not archetypal components of SER systems [2] [3]. The capability to evaluate the reliability and confidence of a classifier’s decision is a crucial aspect in all machine learning problems. This is evident from the abundance of research in the area that has been exerted into ASR systems. Our research aims to extend this notion of confidence to multiclass SER systems. Before describing our approach, we will take a look at the previous work that has been done in extracting confidence from ASR algorithms.

Jiang et al. (2005) have shown that all methods of computing confidence measures in speech recognition can be assigned into three groups [4]. In the first of these groups the aim is to calculate confidence measures by combining predictor features for better performance i.e. n-best lists, log likelihood ratios, acoustic stability and so on. These features are collected during the decoding procedure and then combined in various ways to generate a single score to indicate correctness of the recognition decision. Combination models include the use of Support Vector Machines [2] [5].

Jiang et al.’s second group quantifies posterior word probabilities to estimate a degree of confidence in a given utterance. It is known that conventional ASR algorithms are customarily architected as a pattern classification problem, using the “Maximum a Posterior” rule to resolve the most likely word sequence that achieves the maximum posterior probability. Posterior probabilities are typically estimated from speech system acyclic graphs or N-best lists and are conveyed in the works of Kemp et al. (1997) and Wessel et al. (2001) [3] [6].

The final grouping as outlined by Jiang et al. uses utterance verification and Hidden Markov Models. This involves the recognition of keyword strings that belong to a defined model and the rejection of ones that do not. In this architecture, a Confidence Measure (CM) can be extracted through the statistical testing of two hypotheses i.e. the null hypotheses (word belongs to defined model) and the alternative hypothesis (it does not) [4]. Rahim et al. (1997) outline how likelihood ratios can then be used to quantify whether an utterance belongs to a model or not and how these ratios can be used as measures of confidence [7].

The solutions described above are focused on the domain of ASR, as are all research endeavors in the area of speech recognition confidence to date - except one. Deng et al. (2012) provide a recent contribution to the speech processing community that discusses confidence extraction in SER systems. They propose a semi-supervised algorithm for confidence measurement in emotions grouped by positive or negative valence (+,-) that iteratively assigns labels and trains a group of classifiers. The agreement of these classifiers are combined to calculate a CM for each instance, assessing the correctness of the decisions of the SER system [1].
Our research deviates from that of Deng et al. (2012) above in three ways 1) by using a unique ensemble architecture designed specifically to extract confidence by fusing probability distributions, 2) by being the first to address the more problematic task of confidence mining in multiclass SER systems, rather than the binary problem of valence and 3) by doing so by using a set of metrics not yet applied to ASR or SER systems. Even though mining for confidence in multiple classes does prove to be a more problematic task than in a binary problem, it does provide a more thorough assessment of the overall confidence in SER systems. This paper provides a description of our architecture and experiments, followed by a comparison of various experimental configurations.

II. METHODOLOGY

A. Databases:

For this study, we recorded a subset of instances from the ISEAR (International Survey on Emotion Antecedents and Reactions) to perform our initial experimentation; see Scherer et al. (1994) [8]. These data consist of 7,666 sentences and snippets in which 1096 participants from 16 countries across five continents completed a questionnaire, regarding real life experiences and reactions under multiple emotional categories. The resulting dataset consisted of acted audio recorded in a studio at 16 kHz for the following six fundamental emotional categories: anger (160), disgust (160), fear (160), happiness (160), sadness (160) and surprise (130).

As proof of concept we also held back a second dataset from our initial experiments for evaluation purposes in order to assess our system’s performance on a completely different dataset. For this second dataset we choose eNTERFACE, a widely studied speech database in the speech processing community – also studied by Deng et al. (2012) (Martin et al., 2005) [9]. This corpus consists of recordings from natives of 14 nations that listened to short stories designed to induce a particular emotion. The resulting dataset consists of induced studio recordings at 16 kHz in the emotional categories to be studied: anger (215), disgust (215), fear (215), happiness (207), sadness (210) and surprise (215).

B. Acoustic Features:

To extract the acoustic features for processing our audio instances we use the extraction engine OpenSMILE pioneered by Eyben et al. (2010) [10]. This engine provides the widely used ‘emobase2010’ configuration which includes 34 Low-Level Descriptors (LLDs) along with their corresponding delta coefficients to which functionals are also applied amounting in 1,582 features in total. The LLD’s and functionals are outlined in the Table 1 and Table 2 below.

<table>
<thead>
<tr>
<th>Table - 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LLD</strong></td>
</tr>
<tr>
<td>Intensity &amp; Loudness</td>
</tr>
<tr>
<td>Pitch</td>
</tr>
<tr>
<td>Voice Quality</td>
</tr>
<tr>
<td>LPC</td>
</tr>
<tr>
<td>Cepstrum</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table - 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Functionals</strong></td>
</tr>
<tr>
<td>Extremes</td>
</tr>
<tr>
<td>Means</td>
</tr>
<tr>
<td>Regression</td>
</tr>
<tr>
<td>Moments</td>
</tr>
<tr>
<td>Percentiles</td>
</tr>
<tr>
<td>Time/Duration</td>
</tr>
</tbody>
</table>

We employ an ensemble of seven classifiers. Six of these are statistical classifiers that require training (3 audio, 3 text). To do so a subset of the appropriate datasets are used for the training of the classifiers and the remainder are used for testing. We applied Principle Component Analysis methods to rank the extracted features and all attributes were also feature scaled by normalization.
C. Novel Hybrid Architecture:

Our emotion recognition system defined below is the first of its type to offer fine-grained multimodal classification of the six fundamental emotional classes. It employs three modes employing the combined techniques of 1) acoustic parameter processing 2) text processing in the form of keyword spotting and weighting, WordNet lexicons, application of heuristic rules, incorporation of common abbreviations and colloquialisms, and finally 3) text processing in the form of statistical analysis. The statistical classifiers used for both audio and text were variations of Naive Bayes, Decision Trees and Support Vector Machines (SVMs) using Sequential Minimal Optimization (SMO). The rule based classifications were implemented based on the algorithms of [11].

Since ASR is not the focus of our research, for experimental purposes we feed audio instances along with the corresponding text. We study the output of these separate classifiers in terms of probability distributions. Delany et al. (2005) have shown how numeric outputs from probabilistic classifiers (although providing sense in their decisions) are ‘estimated’ probabilities and not a measure that articulate confidence [12]. Using probabilities however allows us to interpret classifier outputs, giving us more information than merely choosing the class with highest probability. In our experiments we investigate various methods of fusion to form inputs to a meta-decision level classifier, i.e. maximum probability, mean probability, the median of probabilities for a given class and finally a Bayesian updating method of fusion. It is from this meta-classifier that we extract our confidence metrics.

D. Confidence Metrics:

Delany et al. (2005) propose a number of metrics that can be used with the k-NN classifier that accurately articulate prediction confidence, this is something the standard distance output fails to achieve. These metrics have yet to be applied to SER problems and for this reason we use a k-NN classifier at our decision level and introduce the metrics (sections 2.3.1 through 2.3.4) of Delany et al. to the SER domain.

1) Average Nearest Unlike Neighbour:
The Average Nearest Unlike Neighbour Index is a measure of how close the first k NUNs are to the target instance t and is given in Equation 1.

\[ \text{AvgNUNInd}\alpha(t,k) = \frac{\sum_{i=1}^{k} \text{IndexOfNUN}(t)}{k} \]  

\( \text{IndexOfNUN}(t) \) is the index of the ith Nearest Unlike Neighbour (NUN) of the target instance t. By index we mean the ordinal ranking of the example in the list of neighbours.

2) Similarity Ratio:
The Similarity Ratio metric computes the ratio of the similarity between the target instance t and its k Nearest Like Neighbours (NLN) to the similarity between the target instance and its k NUNs. It is given in Equation 2.

\[ \text{SimRatio}(t,k) = \frac{\sum_{t=1}^{k} \text{Sim}(t,\text{NLN}(t)) + e}{\sum_{t=1}^{k} \text{Sim}(t,\text{NUN}(t)) + e} \]  

\( \text{Sim}(a, b) \) in this metric is the calculated similarity between instances \( a \) and \( b \). Here \( e \) is a smoothing value to account for examples that have no NLNs or NUNs (\( e = 0.001 \) was used in all of our calculations)

3) Similarity Ratio Within K:
The Similarity Ratio Within K is similar to the Similarity Ratio as described above except that, it uses only the NLNs and NUNs from the first \( k \) neighbours. It is defined in Equation 3.

\[ \text{SimRatioK}(t,k) = \frac{\sum_{t=1}^{k} \text{Sim}(t,\text{NNN}(t)) + e}{\sum_{t=1}^{k} \text{Sim}(t,\text{NN}(t))(1 - \delta_{ab})} \]  

Here \( \text{Sim}(a, b) \) is the same as above and \( \delta_{ab} \) is Kronecker’s delta where \( \delta_{ab} = 1 \) if the class of a is the same as the class of b and 0 otherwise. Again \( e \) is a smoothing value to account for situations that may have no NUNs among its \( k \) nearest neighbours (\( e = 0.001 \)).

For more information on these k-NN confidence metrics see [12]. Each metric is designed to increase as confidence increases, therefore they are ideal for their intended use. In order to evaluate each measure it is necessary to define a "confidence threshold" i.e. declare that any measure that is greater than the defined threshold is deemed a "confident" instance. We must find an appropriate threshold for each measure in each class. To do so each measure is evaluated from \( k = 1 \) to \( k = 20 \) to identify a threshold for each class, which gives the highest proportion of confident instances while reporting zero confident yet incorrect predictions.
In extracting confidence information from any classifier both Delany et al. (2005) and Deng et al. (2005) agree that a combination approach is generally best in describing the reliability of a prediction using individual components [7][12]. Both argue that this approach can greatly improve the overall performance of the metrics in measuring confidence. Delany et al. (2005) experimentally show how none of the individual measures are consistently effective at predicting confidence, and they therefore define an accumulation approach [12].

4) Accumulated Confidence Measure:
Delany et al.’s Accumulated Confidence Metric (ACM) is determined by assigning confidence to a prediction if any of the three individual measures reveal confidence. In order to achieve this, calculation of the individual thresholds per class must be computed on training data in a pre-classification stage and then the ACM is determined during classification.

III. Experimentation & Results

Our experiments aim to compare the results of a variety of data fusion techniques in terms of both accuracy and confident results. By fusing the probability distributions of our base classifiers we are creating instances for our k-NN meta-classifier. Working on our training data our thresholds per class are programatically determined for each k ∈ {1...20} searching for the value of k that maximizes confidence while having zero confident false positives. Once training is complete and the optimum ks and thresholds that maximize confidence are determined, we run our test cases and declare a confident result if any of the three metrics indicate confidence. Figure 1 below illustrates how each metric performs on the ISEAR dataset when we fuse probabilities by their mean. The linear nature of the ACM reveals how confidence is declared if any of the metrics reveal confidence. We can therefore see that for any value of k at least one of the metrics reveal confidence for 83.17% to 83.33% of instances.

![Fig. 1: Performance of All Confidence Metrics When We Use a Mean/Average To Fuse Base Classifier Probabilities](image)

<table>
<thead>
<tr>
<th>Fusion</th>
<th>Accuracy</th>
<th>Blunders</th>
<th>Confidence</th>
<th>Confidence Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>93.83%</td>
<td>2.67%</td>
<td>83.33%</td>
<td>97.60%</td>
</tr>
<tr>
<td>Max</td>
<td>78.66%</td>
<td>4.33%</td>
<td>64.67%</td>
<td>93.56%</td>
</tr>
<tr>
<td>Median</td>
<td>92.00%</td>
<td>2.83%</td>
<td>77.33%</td>
<td>97.63%</td>
</tr>
<tr>
<td>Bayesian</td>
<td>92.83%</td>
<td>2.00%</td>
<td>81.33%</td>
<td>98.57%</td>
</tr>
<tr>
<td>Fusion</td>
<td>Accuracy</td>
<td>Blunders</td>
<td>Confidence</td>
<td>Confidence Accuracy</td>
</tr>
<tr>
<td>Mean</td>
<td>91.00%</td>
<td>1.83%</td>
<td>85.83%</td>
<td>98.64%</td>
</tr>
<tr>
<td>Max</td>
<td>78.50%</td>
<td>2.83%</td>
<td>65.67%</td>
<td>94.16%</td>
</tr>
<tr>
<td>Median</td>
<td>88.83%</td>
<td>2.67%</td>
<td>78.17%</td>
<td>98.72%</td>
</tr>
<tr>
<td>Bayesian</td>
<td>90.66%</td>
<td>1.50%</td>
<td>79.67%</td>
<td>99.58%</td>
</tr>
</tbody>
</table>

Table 3 above summarizes the performance four prominent data fusion methods used at the meta-classifier level in terms of prediction accuracy, the amount of blunders made and prediction confidence. Prediction accuracies recorded here are the most optimum accuracies recorded for the value of k yielding the highest confidence, therefore optimizing both confidence and accuracy. Here we define classification blunders as those misclassifications that are entirely dissimilar/conflicting to the correct class e.g. happiness in place of anger, disgust, fear, sadness or vice versa. The “Confidence Accuracy” column here highlights how accurately the confidence tagging mechanism performs. In examining these values the data tell us that the metrics employed prove to be extremely efficient in tagging speech emotion data showing as low as 0.42% error using Bayesian fusion in the eNTERFACE dataset (except for when we fuse by maximum probability which shows up to 6.44% error in the ISEAR dataset). We note that this is the poorest method of fusion (the maximum) also yielding the lowest prediction accuracy.

Our results indicate that fusing probability distributions by their mean yields the best classification results in terms of both accuracy and confidence. Note here that in a similar fashion to the first dataset, our evaluation dataset (eNTERFACE) revealed similar results for each of the fusion methods where fusion by mean again proved to be the most efficient method of fusion for
probabilities showing the highest classification confidence of 85.33%. We also note that the Bayesian fusion method performs as good as the mean approach. Even though the reported confidence is slightly less we can see that the reported confidence error is also less. It is therefore beneficial to take into account classifier accuracy, confidence accuracy and confidence error in selecting an optimum SER architecture.

In order to convey confidence as a single number metric using this setup we propose Equation (4) which scales a given confidence metric $x_i$ based on the maximum recorded value for that metric and the defined confidence threshold for the predicted class. The $T$ value here refers to the number of confidence metrics used. Using this technique we managed to achieve up to 0.8146 confidence for unseen test instances in the eINTERFACE evaluation set – an improvement on the 0.7950 reported by Deng et al. (2012) in their binary approach using this set.

$$\text{ACM} = \frac{\sum_{i=1}^{T} x_i - \text{thres}_{si}}{T \cdot (\max x_i) - T \cdot (\min x_i)}$$

Finally it can be seen in Table 3 above that each of the methods of fusion specifically the mean, median and Bayesian methods have a high confidence accuracy, that is to say any time confidence is assigned to a prediction it is correct on average 97.30% of the time across both datasets studied. This table also shows an average of 2.58% of blunderous predictions across both sets. Given this very high confidence accuracy rate and the very low blunder rate our results did not reveal one blunderous prediction that was assigned confidence. These results therefore indicate that there is credence in using a confidence component to serve as a worthwhile aid in providing further information blunderous predictions in SER systems.

IV. CONCLUSIONS

We have described the first successful demonstration of confidence extraction in the multiclass problem of SER. By treating the problem as multiclass rather than binary we are able to show a more fine grained result and therefore a more detailed understanding of both classification and confidence in SER systems. We have also experimentally optimized this system so that it yields both the best possible accuracy and the best possible confidence in its predictions. Furthermore we have demonstrated how our system has shown improvement on a previous binary approach and finally we note how confidence measures can be useful as a tool to catch blunderous predictions in SER problems. This research achieved the above by adopting metrics which are original to the fields of both ASR and SER and showed how useful these metrics can be, hopefully inspiring the community to apply them elsewhere.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the advice of both Michael Madden and Colm O Riordan of the Discipline of Information Technology in the National University of Ireland Galway in carrying out this research.

REFERENCES