LATE FUSION OF INDIVIDUAL ENGINES FOR IMPROVED RECOGNITION OF NEGATIVE EMOTION IN SPEECH – LEARNING VS. DEMOCRATIC VOTE

Björn Schuller[1], *Florian Metze*[2], *Stefan Steidl*[3], *Anton Batliner*[3], *Florian Eyben*[1], *Tim Polzehl*[4]

[1]Institute for Human-Machine Communication, Technische Universität München (TUM), Germany
[2]Language Technologies Institute, Carnegie Mellon University (CMU), Pittsburgh, PA, USA
[3]Lehrstuhl für Mustererkennung, Friedrich-Alexander-Universität (FAU), Erlangen-Nürnberg, Germany
[4]Usability and Quality Lab, Technische Universität Berlin (T-Labs), Berlin, Germany

schuller@tum.de

ABSTRACT

The fusion of multiple recognition engines is known to be able to outperform individual ones, given sufficient independence of methods, models, and knowledge sources. We therefore investigate late fusion of different speech-based recognizers of emotion. Two generally different streams of information are considered: acoustics and linguistics fed by state-of-the-art automatic speech recognition. A total of five emotion recognition engines from different sites that provide heterogeneous output information are integrated by either simple democratic vote or learning 'which predictor to trust when'. We are able to significantly outperform the best individual engine by fusion, and the so far best reported result on the recently introduced Emotion Challenge task.

Index Terms— Emotion Recognition, Late Fusion, Speech Analysis

1. INTRODUCTION

Emotion recognition is notoriously difficult. The 'traditional' way of obtaining good classification performance has been the use of acted, tightly controlled data, or the use of only a tightly controlled, preselected subset of 'prototypical' cases. In the last years, researchers started using more or less realistic databases but still confined analyses onto rather clear cases, by that not modelling rest or garbage classes. Better classifier performance was aimed at by employing a higher number and a greater diversity of acoustic and/or linguistic features, and by employing more sophisticated classifiers. In [1] we could show that a further improvement can be obtained if results obtained with different classifiers and different feature types are combined in a rover approach; still, clear cases were preselected, and 100 % correct speech recognition was assumed by employing the spoken word chain. In the present paper, we are using the same database, but all cases in an 'open microphone' setting; this clearly will yield lower classifier performance because many unclear, non-prototypical cases have to be processed. Moreover, we do not use any longer the spoken word chain but the output of automatic speech recognition, by that coming close to a 'real' processing of emotions 'in the wild'. For the combination we select a late fusion, which most flexibly allows to combine individual emotion recognition engines, and evaluate whether under these more realistic conditions a benefit can be measured. Two variants are investigated: once a simple democratic vote, and once by profiting from output certainty information of all engines. As such tends to be heterogeneous in terms of kind of certainty measure and amount of information provided, a classifier based fusion seems the reasonable choice, which at the same time is able to learn certain confusion or disagreement patterns among the recognition instances.

The remainder of this paper is structured as follows: Sec. 2 introduces the dataset, Sec. 3 5 different emotion recognition engines which are fused in two different late fusion manners within Sec. 4 before results are presented and discussed in Sec. 5 and Sec. 6.

The research leading to these results has received funding from the European Community under grant No. IST-2001-37599 (PF-STAR), grant No. IST-2002-50742 (HUMAINE), and grant (FP7/2007-2013) No. 211486 (SEMAINE). The responsibility lies with the authors.

2. SPEECH DATABASE

The FAU Aibo Emotion Corpus comprises recordings of German children's interactions with Sony's pet robot Aibo; the speech data are spontaneous and emotionally coloured. The children were led to believe that the Aibo was responding to their commands, whereas the robot was actually controlled by a human operator. The wizard caused the Aibo to perform a fixed, predetermined sequence of actions; sometimes the Aibo behaved disobediently, thereby provoking emotional reactions. The data was collected at two different schools, MONT and OHM, from 51 children (age 10 - 13, 21 male, 30 female; about 8.9 hours of speech without pauses). Speech was transmitted with a high quality wireless head set and recorded with a DAT-recorder (16 bit, 48 kHz down-sampled to 16 kHz). The recordings were segmented automatically into 'turns' using a pause threshold of 1 s. 5 labelers listened to the turns in sequential order and annotated each word independently from each other as neutral (default) or as belonging to one of ten other classes. We resort to majority voting (MV): if three or more labelers agreed, the label was attributed to the word. In the following, the number of cases with MV is given in parentheses: *joyful* (101), *surprised* (0), *emphatic* (2 528), *helpless* (3), *touchy*, i. e. irritated (225), *angry* (84), *motherese* (1 260), *bored* (11), *reprimanding* (310), *rest*, i. e. non-neutral, but not belonging to the other categories (3), *neutral* (39 169); 4 707 words had no MV; all in all, there were 48 401 words. Manually defined chunks based on syntactic-prosodic criteria [2, Chap. 5.3.5] are used here. In contrast to other publications published recently, the whole corpus consisting of 18 216 chunks is used under the very same conditions as for the INTERSPEECH Emotion Challenge [3]. In this paper, we concentrate on the two-class problem consisting of the cover classes **NEG**ative (subsuming *angry*, *touchy*, *reprimanding*, and *emphatic*) and **IDL**e (consisting of all non-negative states); note

#	NEG	IDL	$\sum$
train	3 358	6 601	9 959
test	2 465	5 792	8 257

Table 1. Number of instances for the two-class task of the FAU Aibo Emotion Corpus

that emphatic has to be conceived as a pre-stage of anger because on the valence dimension, it lies between neutral and anger, cf. [2]. A heuristic approach similar to the one applied in [2, Chap. 5.3.8] is used to map the raw labels of the 5 labelers on the word level onto one label for the whole chunk: If 50 % of these raw labels are **NEG**, then the whole chunk is labelled as **NEG**. Furthermore, the whole chunk is considered to be **NEG** as well if the following two conditions are fulfilled: 1) at least one third of all raw labels is **NEG**, and 2) the remaining raw labels are mostly pure *neutral*, i. e. at least 90 % of all raw labels are either negative (*angry*, *touchy*, *reprimanding*, *emphatic*) or *neutral*. Frequencies are given in Table 1. Speaker independence is guaranteed by using the data of one school (OHM, 13 male, 13 female) for training and the data of the other school (MONT, 8 male, 17 female) for testing.

3. INDIVIDUAL ENGINES

A number of 5 individual engines is considered as detailed in the following. Each engine is indexed and named by the contributing site (speech recognizer first in case of two sites) and type of feature information used. At the end of each engine description the output information for the latter fusion is detailed.

3.1. Engine 1 - CMU/TUM Linguistic

The CMU Automatic Speech Recognition (ASR) system was trained on about 14 h of close-talking, clean 16 kHz 'background' speech, recorded from adults reading German newspaper texts, using the Janus toolkit and the Ibis decoder. The acoustic model uses 2 k context-dependent, speaker-independent acoustic models. These were trained using Maximum Likelihood (ML) and employ 32 Gaussians with diagonal covariance matrices each in a 42-dimensional MFCC-based feature space after LDA, also using VTLN and speaker-based CMN/CVN. The baseline language model was also trained using tri-grams on German Broadcast News type text data and transcripts, using a 60 k vocabulary. To adapt this system to the target task, we reduced the vocabulary of the original system to 5 k words, 4.5 k of them unique. This includes 300 new domain-specific words appearing at least two times, including non-standard speech, as long as they appeared to be emotionally salient, for which pronunciations were generated manually. We merged the ML update statistics on the 'background' database with matching statistics collected on the training data, using fixed weights, to derive MAP adapted acoustic models. The language model (LM) was also adapted to the target domain using a context independent, LOO-aware interpolation of 3-gram background and in-domain LMs for development. Averaged perplexity on the training data is 55. It is interesting to note that the higher level education school's children comprised in the training partition have a higher vocabulary of 703/253 words/fragments as opposed to the test set's vocabulary size at 383/158. During tests, the baseline acoustic model was adapted to the test speaker incrementally using unsupervised constrained MLLR in the feature space, and VTLN. Speaker adaptation was performed using automatically determined speaker clusters. Trained on the train set, a word accuracy of 81.0 % is obtained for the test set. TUM linguistic emotion analysis is next based on vector space representation: in analogy to bag of words, the bag of n-grams approach also represents text in a numeric feature space. The main difference is the observation of a series of consecutive words as semantic units of interest [4]. The approach allows to observe several n-grams together, determined by a minimum and a maximum n-gram length, similar to 'backing-off'. We found that term frequency, inverse document frequency, chunk length, binary, and case lowering transformations had no influence for the corpus at hand: the average length of a chunk in terms of the number of words is as low as 2.66; **IDL** chunks are 2.82 words long on average, **NEG** chunks only 2.30 words. Only little influence of stemming was observed, which is why it is not used in the experiments. An optimum was further found for the n-gram length of 1 to 3 words. The classifier of choice for these features is a discriminatively learned simple Bayesian Network, namely Discriminative Multinomial Naive Bayes (DMNB). The reason is two-fold: first, the mean recall values resulted in a slight absolute improvement over Support Vector Machines (SVM) as used for our former baseline provision on the FAU Aibo Emotion Corpus [3]: an improvement of 2.05 % / -0.02 % for the described linguistic features (unweighted/weighted average recall). At the same time, DMNB requires lower memory and only a fraction of the computation time of SVM – Sequential Minimal Optimisation training of SVM with linear kernel demanded 200 times higher computation time than DMNB in parameterisation as below using [5] on an 8 GB RAM, 2.4 GHz, 64 Bit industry computer. Second, the parameter learning is carried out by discriminative frequency estimation, whereby the likelihood information and the prediction error are considered. Thus, a combination of generative and discriminative learning is employed. This method is known to work well in highly correlated spaces (as in our case), to converge quickly, and not to suffer from over-fitting. For optimal results we found it best to ignore the frequency information in the data and select a number of only 1 iteration for the linguistic processing. Numeric variables are discretized using unsupervised ten-bin discretization [5]. To ensure that the linguistic emotion model is trained on the same type of phenomena, i. e. ASR errors, it has to face when dealing with the ASR output of the test set, we trained and subsequently tested the ASR engine on the train set. This ASR output of the train set is used for the training of the linguistic emotion model.

This engine delivers the overall score of the chunk (i. e. sentence) hypothesis, the assigned class index together with confidence scores for each class.

3.2. Engine 2 - FAU/TUM Linguistic

To obtain a secondary independent recognition of the spoken word chain from the speech signal, we now use the ASR engine that has been developed within the speech group at the University Erlangen-Nuremberg (FAU). A recent overview is given in [6]. The acoustic features are the first 12 standard MFCC features (the first MFCC coefficient is replaced by the sum of the energies of the 22 Mel filterbanks), and their first derivatives. The features are computed every 10 ms over a Hamming window of 16 ms. This ASR system is based on semi-continuous hidden Markov models (SC-HMM) modelling polyphones, i. e. an extension of the well-known triphones to model large context sizes. A polyphone is modelled by its own HMM if it can be observed at least 50 times in the training set. All HMM states share the same set of Gaussian densities; the size of the codebook is 500. By that, a smaller number of densities can be used,

which is beneficial if – as in our case – only limited training data is available. Yet, full covariance matrices are used in contrast to most systems based on continuous HMMs. Baum-Welch re-estimation for training and Viterbi decoding are used. As language model we use back-off bi-grams. The vocabulary of the ASR system consists of all words (but no word fragments) of both the training and the test set; all in all 813 words. Hence, 158 vocabulary words (types) of the test set are out of vocabulary (OOV), which amounts to a total of 2.1 % OOV events (tokens). This ASR engine, trained on the train set, yields a word accuracy of 77.5 % for the test set. Based on this ASR output linguistic analysis is carried out in full accordance as before.

As output the class index and probability per class are provided.

3.3. Engine 3 - TUM Dynamic Acoustic

This engine was used for the baseline computation in [3]. It employs the low-level-descriptors zero-crossing-rate (ZCR), root mean square (RMS), F0, harmonics-to-noise ratio (HNR), and MFCC 1–12. Classification is realised by continuous density linear left-right HMM (one model per emotion), 5 states with 2 Gaussian mixtures, each, 6+4 Baum-Welch re-estimation iterations, and Viterbi decoding. This parameterisation resembles the optimum found in [3].

This engine outputs the winning class's index together with its log-likelihood score.

3.4. Engine 4 - T-Labs Static Acoustic-Linguistic

The T-Labs prosodic and acoustic system provides a broad variety of information about vocal expression patterns that can be useful when classifying speech metadata in general. Measurements related to voicing such as pitch, zero-crossing rates, and the harmonic-to-noise ratio are extracted. Further, durational information from the patterns of the voiced segments are calculated. Using these patterns, the speech chunks are sub-segmented into voiced, unvoiced, and silence segments, all of which serve for later statistical analysis both separately and jointly. Further processing the time signal, the magnitude of intensity is estimated and its correlation to pitch calculated as separate feature. However, the majority of the systems' features capture spectral information using different approaches. Formant frequencies and the corresponding bandwidths are determined. BARC and MEL filtering are applied in order to produce coefficients of perceptual loudness and MFCCs. In addition, other spectral characteristics like the roll-off point, flux, and centroid are added which are estimated from the spectrum directly without filtering. All these measurements are then taken to a statistical unit that predominantly derives moments, extrema, ranges, regression coefficients, and distributional descriptions. DCT is applied to pitch, intensity, and perceptive loudness directly in order to better capture their dynamics over time. The relative signal power is also measured by building ratios of voiced and unvoiced segments. Finally, voice quality estimation is done by filtering prominent pitch periods followed by pitch-synchronous DFT. After adding delta and double delta coefficients roughly 1.5 k features are obtained which are then taken to an information-gain-ratio based ranking unit. Feature selection on the training set resulted in an optimal feature set size of roughly 320, most of which again belong to features on spectral information. For more details see [7].

The T-Labs linguistic system performs a word recognition task to obtain word hypotheses. For acoustic modeling 2 k context-dependent speaker-independent models are ML trained with 32 Gaussians and diagonal covariance matrices in a 42-dimensional MFCC-based space after LDA as for engine 1. Doing normalization, VTLN and speaker-based CMN/CVN are applied. The language model includes about 5 k words and applies tri-gram modelling. For development on the training data, we computed speaker-specific models and evaluated them in a leave-one-speaker-out manner. As the test data did not provide speaker labels, we did not use a speaker-adaptive ASR system for testing. We however experimented with a speaker-adaptive system that estimated CMN/CVN, VTLN, and constrained MLLR incrementally over a whole speaker which led to a decrease of WER to less than 20 %. To identify emotionally salient words in the utterances, the information-theoretic concept of 'salience' is adopted [8]. The emotional salience of a word for an emotion category is defined as self-mutual information between a specific word and emotion class. The engine calculates the score of emotional salience for each word and aggregates a decision on chunk level. The impact of erroneous word recognition proved marginal.

In terms of classification SVM are used with an RBF kernel for processing the acoustic feature set. The linguistic salience features are subjected to a maximum criterion. After calculating and normalizing confidence scores of both systems, it is finally decided for the classifier offering the highest confidence.

Apart from the winning class's index, the probability per class is given by this engine.

3.5. Engine 5 - TUM Static Acoustic

We limit to a systematic generation of features using our open source feature extraction[1] [9]. In detail, the slightly extended set in comparison to [3] comprises of 26 low-level descriptors: dc offset, extremes, and ZCR from the time signal, RMS and logarithmic frame energy, pitch (F0, normalised to 500 Hz), strength, and quality as well as HNR by autocorrelation function, and MFCC 0–15 in full accordance to HTK-based computation. To each of these, the delta and double delta coefficients are additionally computed. Next the 21 functionals mean, absolute mean, standard deviation, variance, kurtosis, skewness, minimum and maximum value, relative position, and range as well as 2 linear and 3 quadratic regression coefficients with their mean absolute and square errors are applied on a per chunk basis. Thus, the total feature vector per chunk contains $26 \cdot 3 \cdot 21 = 1\,638$ attributes. More details on feature implementation are found in [9]. As for engines 1 and 2, DMNB is used for classification. An improvement of 1.90 % / 2.01 % (unweighted/weighted average recall) for these acoustic features can be named over the usage of SVM. Optimal results were found with 10 iterations for the acoustic processing. As before, the index of the winning class and the prior probabilities per class are output.

4. LATE FUSION

In late fusion architectures signals are integrated at a semantic level. Signals are modeled separately and combined later, during the decoding phase. Each mode has an individual recognizer which is trained independently, so there is no explicit learning of the joint probability of the modalities. Late fusion uses training data from one stream (e. g. acoustic or linguistic – a large number of corpora indeed provides only one such, as many databases either possess only scripted text or are too small to build reliable language models), which is not as rare as such from multiple streams needed

[1] http://sourceforge.net/projects/openSMILE

for early fusion [10] and profits from mature, well-engineered unimodal recognition techniques. Furthermore, late fusion systems scale up easier because no re-training is necessary if further streams or modalities are to be integrated, handling of streams which are temporarily missing (e. g. if no word is output from ASR) is easier, and they provide a higher degree of modularity. However, mutual information coming from another modality is not considered during the recognition of a single mode, which may cause a down-grade in comparison to early fusion if the streams or modalities are correlated as usually in emotion recognition.

Two generally different types of late fusion are considered: a rather simplistic by 'democratic' majority vote, and a meta-classifier-based that may benefit from additional confidence information and learn typical patterns of disagreement among individual recognition engines. In the first case we make exclusively use of the winning classes's index information provided by each of the engines. In the second case we use all of the named output information including different representation forms of class prediction reliability as log-likelihood scores, or actual and pseudo-probabilities of either the speech or emotion recognition or both.

5. EXPERIMENTAL RESULTS

As classes are unbalanced, the primary measure to optimise will be unweighted average (UA) recall, and secondly the weighted average (WA) recall (i. e. accuracy). Constantly picking the majority class would result in an accuracy (WA recall) of 70.1 % for the considered two-class problem, while the chance level for UA recall is 50 %.

Balancing of the training material by random up-sampling to reach equal distribution is used by engine 5 to avoid classifier overfitting. Note that this does not have any influence in the case of dynamic modelling (engine 3): for each class one HMM is trained individually and equal priors are assumed. In the case of linguistic features (all other engines) this was found to result in a down-grade and was thus avoided. Further standardisation of the whole sets, individually, is used by engine 5 to cope with biases due to different room acoustics, etc.

For late fusion by learning we consider three variants which all combine all outputs of all engines as described in Sec. 3: first, without any further processing; second by discretization based on the training set using Kononenko discretization; third by additionally adding a feature that adds all individual indices (NEG:1, IDL:0). A slight improvement can be seen by going from the first to the third variant. Classification is thereby always carried out by DMNB.

Table 2 depicts these results for our two-class tasks. As can be seen, the late fusions significantly (level 0.002, one-tailed test) outperform the best individual engine. While learning provides the overall best result, this is not significantly better than democratic vote among the engines. In a similar manner discretization of the heterogeneous scores and the addition of added class indices helped improve the learning approach, yet not significantly.

Recall [%]	**UA**	**WA**
Engine 1: CMU/TUM linguistic	64.79	60.58
Engine 2: FAU/TUM linguistic	66.05	67.87
Engine 3: TUM dynamic acoustic	66.10	65.24
Engine 4: T-Labs static acoustic-linguistic	68.13	73.25
Engine 5: TUM static acoustic	68.26	65.97
Late fusion (democratic vote)	**70.35**	71.07
Learned fusion	69.90	70.17
Learned fusion (Kononenko)	70.40	71.79
Learned fusion (Kononenko, added sum)	**70.45**	71.62

Table 2. Results on the test partition for the different individual engines and their fusion by either democratic vote or a learned meta-classifier in terms of (un-)weighted average recall (UA/WA).

6. CONCLUSIONS

In this work we have shown how multiple emotion recognizers in a late can fusion significantly outperform the best considered single recognizer in terms of unweighted average recall (UA). Using diverse additional confidence measures in a learning approach was not found significantly better than simple democratic vote. This seems interesting, as the latter can be installed easily and already leads to improvements. Acoustic features produced better results in the considered settings on our FAU Aibo Emotion Corpus. Overall, the best result on this corpus and task obtained so far (70.29 % UA, 68.68 % WA recall) [11] could be outperformed in terms of numbers – significantly for WA, yet not significantly for UA. We next aim at learning of individual labelers and their fusion as expert voters.

7. REFERENCES

[1] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "Combining Efforts for Improving Automatic Classification of Emotional User States," in *Proceedings of IS-LTC 2006*, Ljubliana, 2006, pp. 240–245.

[2] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*, Logos Verlag, Berlin, 2009.

[3] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. Interspeech*, Brighton, UK, 2009.

[4] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Emotion Recognition from Speech: Putting ASR in the Loop," in *Proc. ICASSP*, Taipei, Taiwan, 2009, pp. 4585–4588, IEEE.

[5] I. H. Witten and E. Frank, *Data mining: Practical machine learning tools and techniques, 2nd Edition*, Morgan Kaufmann, San Francisco, 2005.

[6] G. Stemmer, *Modeling Variability in Speech Recognition*, Logos Verlag, Berlin, 2005.

[7] T. Polzehl, S. Sundaram, H. Ketabdar, M. Wagner, and F. Metze, "Emotion Classification in Children's Speech Using Fusion Acoustic and Linguistic Features," in *Proc. Interspeech*, Brighton, UK, 2009.

[8] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

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

[10] W. Lizhong, S. Oviatt, and P. R. Cohen, "Multimodal integration - a statistical view," in *IEEE Transactions on Multimedia*, 1999, vol. 1, pp. 334–341.

[11] P. Dumouchel, N. Dehak, Y. Attabi, R. Dehak, and N. Boufaden, "Cepstral and Long-Term Features for Emotion Recognition," in *Proc. Interspeech*, Brighton, UK, 2009.