Cross-Corpus Data Augmentation for Acoustic Addressee Detection

Oleg Akhtiamov[1,2], **Ingo Siegert**[3], **Alexey Karpov**[4], **Wolfgang Minker**[1]
[1]Ulm University, Ulm, Germany
[2]ITMO University, St. Petersburg, Russia
[3]Otto-von-Guericke-University, Magdeburg, Germany
[4]SPIIRAS, St. Petersburg, Russia
oakhtiamov@gmail.com, ingo.siegert@ovgu.de,
karpov@iias.spb.su, wolfgang.minker@uni-ulm.de

Abstract

Acoustic addressee detection (AD) is a modern paralinguistic and dialogue challenge that especially arises in voice assistants. In the present study, we distinguish addressees in two settings (a conversation between several people and a spoken dialogue system, and a conversation between several adults and a child) and introduce the first competitive baseline (unweighted average recall equals 0.891) for the Voice Assistant Conversation Corpus that models the first setting. We jointly solve both classification problems, using three models: a linear support vector machine dealing with acoustic functionals and two neural networks utilising raw waveforms alongside with acoustic low-level descriptors. We investigate how different corpora influence each other, applying the mixup approach to data augmentation. We also study the influence of various acoustic context lengths on AD. Two-second speech fragments turn out to be sufficient for reliable AD. Mixup is shown to be beneficial for merging acoustic data (extracted features but not raw waveforms) from different domains that allows us to reach a higher classification performance on human-machine AD and also for training a multipurpose neural network that is capable of solving both human-machine and adult-child AD problems.

1 Introduction

For the past years, the phenomenon of multiparty spoken interaction has drawn many researchers' attention (Busso et al., 2007; Gilmartin et al., 2018; Haider et al., 2018). How do we address other people in such conversations? Normally, we do this either explicitly, directly specifying desirable addressees by their names, or implicitly, using contextual (Ouchi and Tsuboi, 2016; Zhang et al., 2018) and multimodal markers (Tsai et al., 2015; Akhtiamov et al., 2017b; Akhtiamov and Palkov, 2018; Le Minh et al., 2018). Particularly, we use acoustic markers to emphasise special addressees, such as hard-of-hearing people (Batliner et al., 2008), elderly people, children (Schuller et al., 2017), and automatic spoken dialogue systems (SDSs) (Batliner et al., 2008; Shriberg et al., 2013; Akhtiamov et al., 2017a; Pugachev et al., 2017). We act in this way if we realise that our addressee may have some communicational difficulties, and therefore we modify our normal manner of speech, making it more rhythmical, louder, and generally more understandable as soon as we start talking to such conversational partners (Shriberg et al., 2012; Siegert and Krüger, 2018).

In the present research, we deal with two binary acoustic addressee detection (AD) problems. The first problem of human-machine addressee detection (H-M AD) arises in conversations within a group of users solving a cooperative task by means of an SDS. The users may talk to each other and also contact the system from time to time. The system is supposed to distinguish between machine- and human-directed utterances in order to maintain conversations in a realistic manner. Human-directed utterances do not require a direct system response and should be processed with the system in an implicit way. We use the following two corpora to model the H-M AD problem: the Smart Video Corpus (SVC) (Batliner et al., 2008) and the Voice Assistant Conversation Corpus (VACC) (Siegert et al., 2018). The first competitive VACC baseline is introduced in the present paper. The second problem of adult-child addressee detection (A-C AD) appears in conversations between a group of adults and a child. In this case, our system is supposed to distinguish between child- and adult-directed utterances. A possible application for such a system of adult-child conversation monitoring is the estimation of children's and adults' conversational behaviour that will allow us

Proceedings of the SIGDial 2019 Conference, pages 274–283
Stockholm, Sweden, 11-13 September 2019. ©2019 Association for Computational Linguistics

to measure Interaction Quality (IQ) (Spirina et al., 2016). According to this complex metric, we will be able to assess the children's progress in maintaining conversations. We model the A-C AD problem, using the HomeBank Child-Adult Addressee Corpus (HB-CHAAC, mentioned as HB below for simplicity) (Casillas et al., 2017).

We consider both binary classification problems as one: the utterances belonging to the first category are directed to a special addressee that may be an SDS or a child having a lack of communicational skills. The utterances belonging to the second category are directed to ordinary adults without any impairments that may cause miscommunication. In this light, we conduct a series of cross-corpus experiments and merge several corpora with the *mixup* method. This data augmentation technique has already been studied on image classification (Zhang et al., 2017), speech recognition (Medennikov et al., 2018), and acoustic emotion recognition (Fedotov et al., 2018b).

The present paper has the following contributions: the H-M and the A-C AD problem are jointly analysed by means of machine learning; mixup in combination with state-of-the-art classifiers is applied to cross-corpus acoustic AD for the first time; mixup capabilities are investigated over various speech signal representations (including raw data), acoustic context lengths, corpora, domains, languages, and classification problems.

2 Related Work

Several studies have already been conducted on the problem of acoustic H-M AD. The current acoustic SVC baseline was introduced by Akhtiamov et al. (2017a), who applied a feature selection method to a large paralinguistic feature set containing various functionals computed over low-level descriptor (LLD) contours (2013 ComParE feature set described by Eyben (2015)). The ComParE LLDs and their functionals were shown to be a universal solution for a wide range of paralinguistic problems besides AD, e.g, acoustic emotion recognition (Fedotov et al., 2018a), native speech detection, and neurological pathology estimation (Schuller et al., 2015). The same attribute set in combination with a linear support vector machine (SVM) alongside with other models including an end-to-end neural network was applied to the problem of acoustic A-C AD on HB by Schuller et al. (2017). HB was introduced within the Addressee Sub-Challenge of the Interspeech 2017 Computational Paralinguistics Challenge (ComParE) (Schuller et al., 2017) that has already been finished. However, the challenge organisers proposed an extremely competitive baseline (Schuller et al., 2017) that none of the challenge participants managed to surpass, and therefore the HB classification problem remains of great scientific and practical interest.

There also exist speech signal representations designed specially for acoustic H-M AD. Shriberg et al. (2013) suggested modelling speech rhythm and vocal effort with high-abstract attributes: energy contour features, voice quality and spectral tilt features, and delta energy at voicing onsets/offsets. The energy contour and tilt features employed Gaussian mixture models (GMMs) to compute a log likelihood ratio of the two addressee classes. The machine-directed utterances from the corpus used for experiments in the latter study were short predefined commands consisting of three words on average. However, the machine- and child-directed utterances from the data that we have at our disposal were recorded under real-life conditions and usually contain whole sentences of spontaneous speech. Furthermore, it is unclear how these specific attributes perform on A-C AD. Therefore, we would not like to confine to such a narrow attribute set. Instead, we want to use the ComParE features in order to capture all the variety of spontaneous speech. An argument in favour of low-level features, such as LLDs and raw data, is the possibility to use them in combination with deep neural networks capable of performing feature selection and feature transformation implicitly for a specific problem. In the present study, we apply the ComParE functionals jointly with simple linear models, while lower-level features (raw audio and the ComParE LLDs) are used in combination with deep neural networks that learn high-level feature representations for our AD problem. Compared to Mallidi et al. (2018), we do not have that much data for training our networks on acoustic AD. We offset this lack by means of data augmentation.

3 Proposed Approach

3.1 Classifiers

We apply the following three models to audio classification. The first classifier (**func**) is a simple SVM with a linear kernel (Hofmann and Klinkenberg, 2013). This model deals with the ComParE feature set comprising 6373 functionals (Eyben, 2015) extracted at the utterance level.

The second classifier (**LLD**) consists of two stacked long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) layers followed by a global max pooling, a dropout (Srivastava et al., 2014), and a softmax layer. As input, the first layer receives the same LLD sequences used for computing the ComParE functionals. Each sequence element is a vector of 130 LLDs extracted for a sliding time window of 60 ms with an overlap of 50 ms. The sequences are extracted from acoustic context windows of various lengths (from 1/8 to 8 s). The context windows are cut out of audio files with an overlap of 75%. The predictions obtained on several windows belonging to one utterance are averaged to get the final utterance-level prediction.

The third classifier (**e2e**) performing end-to-end speech signal processing differs from the second model in the following way: the sequences of the ComParE LLDs are replaced by the output of a convolutional neural network (CNN). As a result, we obtain a convolutional recurrent neural network (CRNN) that is quite similar to the one suggested by Trigeorgis et al. (2016) for acoustic emotion recognition. However, the initial network architecture specified in the latter study did not provide any reliable results on our AD problem probably due to a lack of perceptive abilities. For this reason, we replaced the initial two-layer CNN by a deeper one. We took the five-layer SoundNet architecture (Aytar et al., 2016) as the reference point for our CNN, cut off its last convolutional layer and scaled the filter sizes and the number of units in each layer in accordance with the input signal resolution and the available amount of our training data. The final shape of the e2e model is depicted in Figure 1.

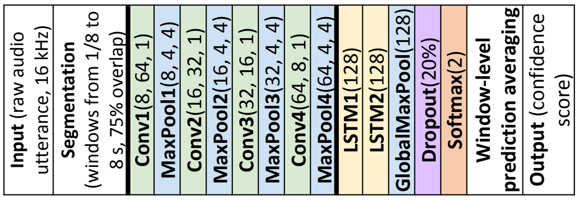


Figure 1: E2e classifier. To obtain the LLD model, we replace the middle part of the e2e model by the ComParE LLD sequences. Notation of the layers in the middle part of the e2e model: *layer_name(n_units, filter_size, stride)*, other layers: *layer_name(n_units)*.

For the func and LLD models, we use statistical corpus normalisation by bringing the handcrafted features to zero mean and unit variance. For the e2e model, we employ batch normalisation (Ioffe and Szegedy, 2015) between each convolution and activation instead. Training our neural networks, we use the following parameters optimised on a development set: Gaussian noise applied to the input signal if mixup is disabled, 20% dropout applied directly before the softmax layer, rectified linear unit (ReLU) as an activation function for all convolutional layers, categorical cross-entropy as a loss function, Adam (Kingma and Ba, 2014) as a weight optimisation algorithm, 100 epochs, and a batch size of 32 examples. The initial learning rate is chosen from the set $\{10^{-3}, 10^{-4}, 10^{-5}\}$ and then divided by 10 if there is no performance improvement observed for the past 10 epochs on the development set. We make checkpoints, saving the current model weights at each epoch and using the best checkpoint as the resulting model according to its performance on the development set.

Both neural networks were designed with TensorFlow (Abadi et al., 2016). The func model was implemented with RapidMiner (Hofmann and Klinkenberg, 2013). We used the openSMILE toolkit (Eyben et al., 2013) and its 2013 ComParE feature configuration (Eyben, 2015) to extract acoustic LLDs and their functionals.

3.2 Data Augmentation

We apply a simple yet efficient approach to data augmentation called *mixup* (Zhang et al., 2017). The core idea of this method is to regularise our model by encouraging it to make linear predictions in the vector space between seen data points. The method generates artificial examples as linear combinations of the feature and label vectors taken from two arbitrary real examples and mixed at a proportion λ in the following way:

$$x_{art} = \lambda x_i + (1 - \lambda) x_j, \qquad (1)$$

$$y_{art} = \lambda y_i + (1 - \lambda) y_j. \qquad (2)$$

λ is randomly generated from a β-distribution for each artificial example. This distribution is defined as follows by a coefficient α that lies within the interval (0, ∞) and determines the probability that our generated example lies close to one of real examples:

$$f(x; \alpha) = x^{\alpha-1}(1 - x)^{\alpha-1}. \qquad (3)$$

VACC (German)				SVC (German)				HB (English)			
Label	Train	Dev	Test	Label	Train	Dev	Test	Label	Train	Dev	Test
M	1809	501	1493	M	546	90	442	C	1882	420	2182
H	862	218	756	H	557	135	423	A	1160	280	1368
Total	2671	719	2249	Total	1103	225	865	Total	3042	700	3550
	(12)	(3)	(10)		(48)	(10)	(41)		(No speaker info)		
	5639 (25), 2:50:20 s				2193 (99), 3:27:35 s				7292, 3:12:16 s		

Table 1: General characteristics of the considered data sets and their utterance-level labelling. Number of speakers is specified in parentheses. Utterance labels: H - human-, M - machine-, A - adult-, C - child-directed. It is assumed that H = A and M = C.

If y_i and y_j from Equation 2 are different hard targets (one-hot vectors) of a classification problem, y_{art} will be a soft target. This solution provides better model regularisation and generalisation over various classes and partially resolves the problem of imbalanced data.

We declare another mixup parameter k that defines the proportion of the number of artificial examples that should be generated and the number of real examples. When merging n corpora, we generate one batch from each corpus, increasing the amount of training data in n times without using mixup. If we simultaneously apply mixup, artificial batches are generated on the fly from n real batches, increasing the amount of training data in $n(k+1)$ times without any considerable delays in the training process. In most of the mixup applications investigated by Zhang et al. (2017), α lies within the interval [0.1, 0.5], i.e., the algorithm biases toward original examples and thereby generates more realistic artificial ones. We use constant α and k values that equal 0.5 and 2 respectively. For greater α values, mixup leads to underfitting.

4 Corpora

We examine our models on the audio data of the three corpora mentioned above. The VACC data set contains experimental conversations in German between a user, a confederate, and an Echo Dot Amazon Alexa device (Siegert et al., 2018). The SVC data set was collected within large-scale Wizard-of-Oz (WOZ) experiments and consists of realistic conversations in German between a user, a confederate, and a mobile SDS (Batliner et al., 2008). For compatibility with the other corpora, we consider the two-class SVC problem introduced by Batliner et al. (2008). The HB data set contains spoken conversations in English between a child and a group of adults recorded under real-life conditions (Casillas et al., 2017). Each corpus was split into a training, a development, and a test set at a proportion defined by its developers. There was no development set specified for SVC by Batliner et al. (2008), and therefore we use 20% of the speakers from its initial training set as a development set. The HB test labels are unavailable to us since this corpus was a part of the Interspeech 2017 ComParE Challenge (Schuller et al., 2017) that has already been finished (none of the participants managed to surpass the Addressee Sub-Challenge baseline). Therefore, we use its development set as a new test set and also utilise 20% of the utterances from its initial training set as a new development set. The partitions of the considered corpora are presented in Table 1. A kernel density estimation (KDE) is depicted in Figure 2 for the utterance length distribution of each corpus.

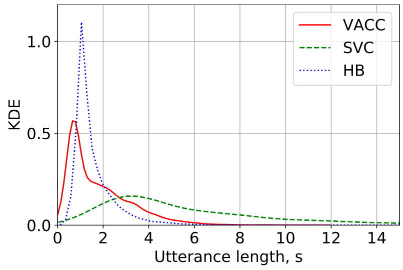
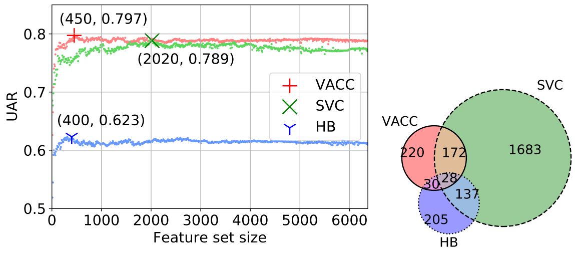


Figure 2: Kernel density estimation (KDE) of the utterance length distributions.

5 Preliminary Experiments with Linear Models

5.1 Feature Selection

Before training the neural network-based classifiers, we conduct preliminary experiments with the func model, aiming to estimate the degree of similarity between the corpora. After feature extraction with the ComParE configuration, we perform recursive feature elimination (RFE), using the coefficients of the normal vector of a linear SVM as attribute weights similarly to Akhtiamov et al. (2017a). Figure 3a demonstrates RFE curves obtained by applying ten-fold leave-one-speaker-group-out cross-validation (LOSGO) on each corpus without its test set. The resulting performance is calculated as unweighted average recall (UAR) for comparability with the existing studies and averaged over all folds for each reduced feature set. A feature set is considered to be optimal if further RFE leads to a stable performance loss. For each corpus, we choose one optimal feature set obtained on a random fold and analyse their intersection depicted in Figure 3b. The representative acoustic attributes vary essentially: VACC, SVC, and HB have only 450, 2020, and 400 relevant features out of 6373 respectively, while having only 28 features in common: some functionals over *F0final sma*, *audSpec Rfilt sma*, *mfcc sma*, *pcm fftMag spectralRollOff25.0 sma*, *pcm fftMag spectralRollOff50.0 sma*, *voicingFinalUnclipped sma*, and their *deltas* (Eyben, 2015). Besides these attributes, VACC and SVC have only 172 features in common, though these two corpora have the same target classes. The optimal feature set size for SVC is considerably greater than for the other two corpora. This difference was probably caused by the WOZ modelling of SVC dialogues as the WOZ setup did not seem convincing enough to some users, resulting in fuzzy addressee patterns that concerned a greater number of acoustic features.

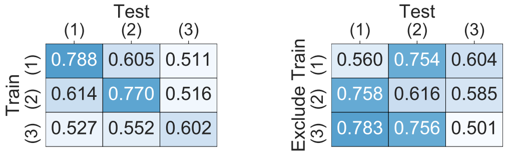


(a) (b)

Figure 3: Preliminary analysis: performance losses during RFE (a), and optimal feature set comparison (b).

5.2 Cross-Corpus and Multitask Classification

We conduct a series of cross-corpus and multitask experiments with the func model, applying a leave-one-corpus-out (LOCO) and an inverse LOCO scheme. In the first scheme, the model is trained on a mixture of all the corpora but one and tested on each of the three corpora. In the second scheme, the model is trained on one corpus and tested on each of the three corpora. In both cases, the model is trained and tested on the corresponding partitions from Table 1. In these experiments, we do not perform feature selection and do not use mixup. Results of the two experimental series are depicted in Figure 4. Let us denote the matrix from Figure 4a as $\bar{A}$, its element as $\bar{a}_{i,j}$, the matrix from Figure 4b as $\bar{B}$, and its element as $\bar{b}_{i,j}$. The resulting UAR ($\bar{a}_{2,2}$) and the optimal feature set size on SVC slightly differ from those obtained by Akhtiamov et al. (2017a) since we apply statistical corpus normalisation in the present study instead of speaker normalisation in order to make our results fairer as the system may face unknown speakers in real applications. Furthermore, there is no information regarding speakers available for HB. $\bar{a}_{1,2}$ and $\bar{a}_{2,1}$ are considerably greater than the other off-diagonal elements of $\bar{A}$, demonstrating a clear relation between VACC and SVC. This result motivates us to explore the potential of the cross-corpus data augmentation on VACC and SVC by means of mixup and deep learning in our future experiments. $\bar{A}$ does not reveal any relation between HB and the other two corpora, though an interesting trend may be noted in $\bar{B}$. $\bar{b}_{2,1}$ and $\bar{b}_{3,1}$ are similar to $\bar{a}_{1,1}$, $\bar{b}_{1,2}$ and $\bar{b}_{3,2}$ are close to $\bar{a}_{2,2}$, and $\bar{b}_{1,3}$ and $\bar{b}_{2,3}$ are similar to $\bar{a}_{3,3}$. Altogether, these three results mean that a single func model trained on examples from two arbitrary corpora demonstrates an adequate performance on them both as if the model were trained on each corpora separately or, in other words, that the three classification problems are non-contradictory. However, A-C AD turned out to be essentially more challenging than H-M AD.



Train \ Test	(1)	(2)	(3)
(1)	0.788	0.605	0.511
(2)	0.614	0.770	0.516
(3)	0.527	0.552	0.602

(a)

Exclude Train \ Test	(1)	(2)	(3)
(1)	0.560	0.754	0.604
(2)	0.758	0.616	0.585
(3)	0.783	0.756	0.501

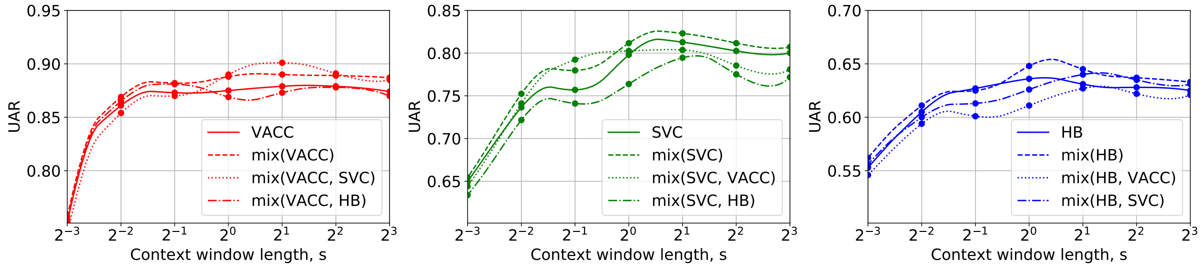
(b)

Figure 4: Results of the inverse LOCO (a) and LOCO (b) experiments with the func model. All values are presented in terms of UAR. Corpora: (1) - VACC, (2) - SVC, (3) - HB.

6 Experiments with Neural Networks

6.1 Mixup and Acoustic Context Length

All the experiments below are presented in terms of UAR for comparability with the existing studies. All statistical comparisons are drawn apply-



(a) Performance on VACC.

(b) Performance on SVC.

(c) Performance on HB.

Figure 5: Classification performance of the LLD model over various context windows and its trends after data augmentation on the considered corpora. In each of the three cases, the training set of the target corpus (on the test set of which UAR is measured) is mixed with itself (*mix(corpus)*) or with itself and with the training set of another corpus (*mix(corpus, another_corpus)*). The points connected with spline interpolation denote exact measurements.

ing a *t*-test with a significance level of 0.05. First, we analyse the sensitivity of our neural networks to acoustic context length variations. This hyperparameter was shown to be critical for paralinguistic problems (Fedotov et al., 2018a). We take a context window length of 1 s as a reference point and then vary it by raising to different powers of two. The context windows are cut out of the audio files with an overlap of 75%. This preprocessing partially resolves the lack of training data. It is possible to align the obtained logarithmic scale with basic acoustic units: given the mean syllable duration estimated by Greenberg (1999) for spontaneous English, we roughly assume that the time intervals between 0, 0.125, 0.5, 1, 2, and 8 s correspond to allophones, syllables, words, collocations/syntagmas, and utterances respectively. In fact, these intervals may significantly overlap since syllable duration is known to be highly speaker-dependent (Greenberg et al., 2003). German words and more complex acoustic units have longer durations compared to their English equivalents.

Performance curves of the LLD classifier tested on LOSGO are depicted in Figure 5. The resulting UAR values are averaged over all folds. The dashed curve is located above the solid one in all three cases, i.e., mixup results in a significant performance improvement already when applied to the same corpus. Adding another corpus to the mixup procedure influences the performance depending on a context window length. Mix(VACC, SVC) significantly surpasses mix(VACC) on VACC for a context window of 2 s. Mix(SVC, VACC) significantly outperforms mix(SVC) on SVC for a context window of 0.5 s. A possible explanation for these two results is that SVC has generally longer utterances (Figure 2) and probably longer acoustic addressee patterns compared to VACC. Mix(HB) does not benefit from adding another corpus to the mixup procedure.

The curves from Figure 5a flatten beyond 0.5 s, meaning that VACC is less sensitive to context length variations than SVC and HB. The optimal context window length, which provides the highest UAR, is 2 s for VACC and SVC and 1 s for HB. However, the latter corpus demonstrates virtually the same result for a longer window of 2 s. The e2e model shows a similar behaviour on various context windows and reaches the highest UAR for the same context window of 2 s on all three corpora. This fact motivates us to confine to a single context window length of 2 s in our future experiments that corresponds to acoustic patterns at the utterance level. Our results confirm an earlier conclusion drawn by Shriberg et al. (2013) regarding the optimal acoustic context length for H-M AD in English.

Table 2 contains the exact UAR values of the two-second performance slices for both neural networks. Similarly to the results presented in Figure 5, the values from Table 2 are obtained on LOSGO and averaged over all folds. The LLD model demonstrates a higher performance com-

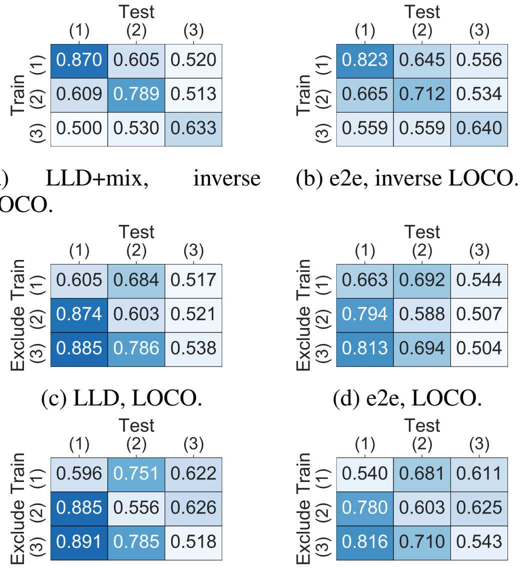
Test Corpus	Model	——	- - -	· · · ·	- · - ·	mix (all)
VACC	LLD	.879	.890	**.901**	.873	.886
	e2e	.853	.834	.852	.845	.846
SVC	LLD	.813	**.823**	.804	.795	.818
	e2e	.764	.756	.758	.749	.761
HB	LLD	.631	.645	.627	.640	.636
	e2e	**.647**	.632	.633	.616	.631

Table 2: Two-second UAR slices. Each marker corresponds to a curve of the same style in Figure 5.

pared to the e2e model overall, except HB, on which both classifiers behave similarly. In contrast to the LLD model, the e2e classifier does not benefit from mixup. This result contradicts the supposition made by Zhang et al. (2017) to apply mixup to raw speech data and may be naturally explained in the following way: after applying mixup to raw speech signals, our augmented data sounds like crowd noise that confuses the e2e model being unable to handle the cocktail party effect. This is not the case for some handcrafted features, e.g., logarithmic attributes, as applying mixup to them does not necessarily mean a simple overlapping of two waveforms, from which these features were extracted. We conclude that applying mixup makes more sense for acoustic features of a higher abstraction level than raw data, e.g., handcrafted LLDs or features extracted with a CNN. In the present study, we confine to two extreme cases: handcrafted LLDs and raw waveforms.

6.2 Cross-Corpus and Multitask Classification

The experiments below are conducted on the partitions specified in Table 1. Six series of cross-corpus experiments are depicted as performance matrices in Figure 6. Let us denote the matrix from Figure 6a as A and its element as $a_{i,j}$, the matrix from Figure 6b as B and its element as $b_{i,j}$, etc. A and B show inverse LOCO experiments on the LLD model with mixup and on the e2e model without mixup respectively. $a_{1,2}$ and $a_{2,1}$ are considerably greater than the other off-diagonal elements of A. $b_{1,2}$ and $b_{2,1}$ are also significantly greater than the other off-diagonal elements of B. Similarly to the matrix $\bar{A}$ from Figure 4a, these two results demonstrate a clear relation between VACC and SVC that was better captured with the e2e model. The other four matrices from Figure 6 contain results of LOCO experiments: C and D - without mixup, E and F - with mixup. The elements $c_{1,3}$, $c_{2,3}$, $d_{1,3}$, and $d_{2,3}$ are close to a random-choice UAR of 0.5, meaning that both neural networks perceive HB as noise and completely ignore it in favour of another corpus. However, the situation changes if we apply mixup: the elements $e_{1,3}$ and $e_{2,3}$ are similar to $a_{3,3}$ as well as the elements $f_{1,3}$ and $f_{2,3}$ being close to $b_{3,3}$. These two results mean that both neural networks start perceiving both corpora involved in the mixup procedure as efficiently as if the networks were trained on each data set separately. Due to a simpler model architecture, the func classifier did not face such a problem of overfitting to a specific corpus during the experiments with multitask learning presented in Figure 4b.



(a) LLD+mix, inverse LOCO.

(b) e2e, inverse LOCO.

(c) LLD, LOCO.

(d) e2e, LOCO.

(e) LLD+mix, LOCO.

(f) e2e+mix, LOCO.

Figure 6: Results of the inverse LOCO and LOCO experiments with the neural networks. All values are presented in terms of UAR. Corpora: (1) - VACC, (2) - SVC, (3) - HB.

A similar trend may be noted in Figure 7 that demonstrates experiments on merging all three corpora: if trained on all the corpora without mixup, both LLD and e2e models discriminate SVC and completely ignore HB. Mixup allows us to train a multipurpose neural network that performs equally well on each of the corpora as if there were three networks trained exclusively for single tasks. The classification performance ob-

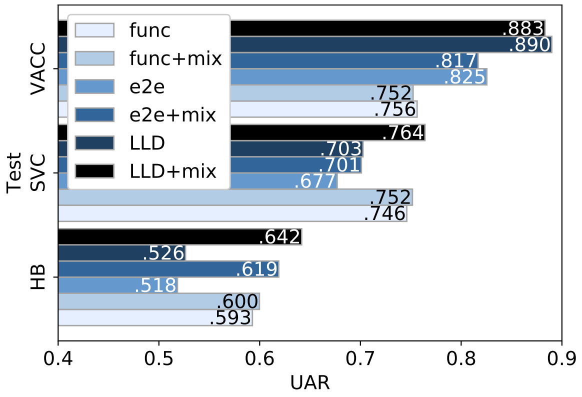


Figure 7: Results of the experiments on merging all three corpora.

tained on VACC and HB with the func model is generally lower compared to the results of the neural networks, and mixup is unable to improve it. However, the func classifier does not suffer from overfitting to a specific corpus during multitask learning and does not need to be regularised.

7 Experiments with ASR-Based Metafeatures

Some metafeatures obtained from an automatic speech recogniser (ASR) are useful for H-M AD since people speak more clearly than usual when addressing an SDS. Machine-directed speech tends to match the ASR patterns better compared to human-directed speech, resulting in a higher ASR confidence (Tsai et al., 2015). It is interesting to check this approach on A-C AD. Using the Google Cloud ASR for German (on VACC and SVC) and for English (on HB), we extract the following ASR metafeatures at the utterance level: *confidence of the best hypothesis*, *number of hypotheses*, *number of words in the best hypothesis*, and *utterance duration in seconds*. These features except the first one (it is already normalised) are brought to zero mean and unit variance and fed to an SVM with a radial kernel (Hofmann and Klinkenberg, 2013). The UAR values obtained with this classifier on the test partitions from Table 1 are equal to 0.778, 0.657, and 0.515 for VACC, SVC, and HB respectively. The latter value is slightly above a random-choice UAR of 0.5, meaning that ASR confidence is non-representative for A-C AD.

8 Conclusions and Future Work

The H-M and A-C AD problems turned out to be essentially different in certain aspects. The first aspect concerns the previously discussed acoustic patterns of child- and machine-directed speech. On the one hand, none of the considered models managed to reveal any relations between HB and the other two corpora during our inverse LOCO experiments. On the other hand, the LOCO experiments with the linear model demonstrate that the H-M and A-C AD problems are non-contradictory. The second aspect is connected with the degree of how often misunderstanding situations occur in an H-M conversation. People tend to talk to the system in a normal manner in the absence of such situations, and this manner of speech may be acoustically undistinguishable from human-directed speech. The third aspect concerns what is said during an A-C conversation. Adults' speech often contains separate sounds and intonations and no verbal information when they talk to children, and therefore ASR confidence is non-representative for A-C AD, though it is useful for H-M AD.

Mixup has been shown to be beneficial for neural networks using predefined acoustic features, while not giving any significant performance improvement for e2e models, though Zhang et al. (2017) supposed that it is worth applying the method to raw speech data as well. Linear classifiers do not benefit from mixup neither due to their simple architectures that do not require any regularisation. Another remarkable capability of mixup was revealed in multitask experiments and applies to both handcrafted features and raw data. This method allows us to merge several corpora modelling similar classification tasks in such a way that one neural network trained on this mixture solves all the tasks equally efficiently with single neural networks, each of which was trained on its own corpus. The corpora being utilised for multitask learning may essentially differ, e.g., VACC and SVC were collected in completely different domains, and HB was even collected for another task and uttered in another language. Without mixup, the neural network overfits to the corpus with the strongest correlation between its features and labels (VACC) and starts discriminating the other corpora. Linear models do not suffer from this problem, though they demonstrate a lower classification performance overall.

Two-second speech fragments are optimal for AD and correspond to acoustic patterns at the utterance level. This result confirms an earlier conclusion drawn by Shriberg et al. (2013) regarding H-M AD in English. According to our inverse LOCO experiments, there exists a clear relation between VACC and SVC. Furthermore, applying mixup to these two corpora allows us to improve classification results on VACC significantly. The following UAR values may be taken from Figure 6 as the new baselines: $e_{3,1} = 0.891$ for VACC and $b_{3,3} = 0.640$ for HB. $b_{3,3}$ is the best baseline for standalone classifiers compared to the results introduced by Schuller et al. (2017) on the original HB development set. Our e2e model surpasses the one from (Schuller et al., 2017) that demonstrated a UAR of 0.609. We achieved this performance improvement due to a more careful choice of the CNN architecture. $a_{2,2} = 0.789$ is similar to the latest SVC baseline of 0.800 established by Akhtiamov et al. (2017a).

In our future work, we plan to extend our experiments, applying mixup to two-dimensional spectrograms and to features extracted with a CNN.

Acknowledgements

The research presented in Section 5 and Section 7 was conducted exclusively within the framework of the Russian Science Foundation project (No. 18-11-00145). The remaining part of the research was supported by DAAD jointly with the Ministry of Science and Higher Education of the Russian Federation within the Michail Lomonosov Program (project No. 8.704.2016/DAAD).

References

Martín Abadi et al. 2016. Tensorflow: A system for large-scale machine learning. In *Symposium on Operating Systems Design and Implementation (OSDI)*, pages 265–283.

Oleg Akhtiamov and Vasily Palkov. 2018. Gaze, prosody and semantics: Relevance of various multimodal signals to addressee detection in human-human-computer conversations. In *International Conference on Speech and Computer (SPECOM)*, pages 1–10. Springer.

Oleg Akhtiamov, Maxim Sidorov, Alexey Karpov, and Wolfgang Minker. 2017a. Speech and text analysis for multimodal addressee detection in human-human-computer interaction. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2521–2525. ISCA.

Oleg Akhtiamov, Dmitrii Ubskii, Evgeniia Feldina, Aleksei Pugachev, Alexey Karpov, and Wolfgang Minker. 2017b. Are you addressing me? Multimodal addressee detection in human-human-computer conversations. In *International Conference on Speech and Computer (SPECOM)*, pages 152–161. Springer.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. SoundNet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems 29*, pages 892–900.

Anton Batliner, Christian Hacker, and Elmar Nöth. 2008. To talk or not to talk with a computer. *Journal on Multimodal User Interfaces*, 2(3):171–186.

Carlos Busso, Panayiotis Georgiou, and Shrikanth Narayanan. 2007. Real-time monitoring of participants' interaction in a meeting using audio-visual sensors. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 685–688. IEEE.

Marisa Casillas, Andrei Amatuni, Amanda Seidl, Melanie Soderstrom, Anne Warlaumont, and Elika Bergelson. 2017. What do babies hear? Analyses of child- and adult-directed speech. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2093–2097. ISCA.

Florian Eyben. 2015. *Real-time speech and music classification by large audio feature space extraction*. Springer.

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *ACM international conference on Multimedia*, pages 835–838. ACM.

Dmitrii Fedotov, Denis Ivanko, Maxim Sidorov, and Wolfgang Minker. 2018a. Contextual dependencies in time-continuous multidimensional affect recognition. In *International Conference on Language Resources and Evaluation (LREC)*, pages 1220–1224. ELRA.

Dmitrii Fedotov, Heysem Kaya, and Alexey Karpov. 2018b. Context modeling for cross-corpus dimensional acoustic emotion recognition: Challenges and mixup. In *International Conference on Speech and Computer (SPECOM)*, pages 155–165. Springer.

Emer Gilmartin, Benjamin R Cowan, Carl Vogel, and Nick Campbell. 2018. Explorations in multiparty casual social talk and its relevance for social human machine dialogue. *Journal on Multimodal User Interfaces*, 12(4):297–308.

Steven Greenberg. 1999. Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29(2-4):159–176.

Steven Greenberg, Hannah Carvey, Leah Hitchcock, and Shuangyu Chang. 2003. Temporal properties of spontaneous speech – A syllable-centric perspective. *Journal of Phonetics*, 31(3):465–485.

Fasih Haider, Hayakawa Akira, Saturnino Luz, Carl Vogel, and Nick Campbell. 2018. On-talk and off-talk detection: A discrete wavelet transform analysis of electroencephalogram. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964. IEEE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Markus Hofmann and Ralf Klinkenberg. 2013. *RapidMiner: Data mining use cases and business analytics applications*. CRC Press.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Thao Le Minh, Nobuyuki Shimizu, Takashi Miyazaki, and Koichi Shinoda. 2018. Deep learning based multi-modal addressee recognition in visual scenes with utterances. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1546–1553.

Sri Harish Mallidi, Roland Maas, Kyle Goehner, Ariya Rastrow, Spyros Matsoukas, and Björn Hoffmeister. 2018. Device-directed utterance detection. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1225–1228. ISCA.

Ivan Medennikov, Yuri Khokhlov, Aleksei Romanenko, Dmitry Popov, Natalia Tomashenko, Ivan Sorokin, and Alexander Zatvornitskiy. 2018. An investigation of mixup training strategies for acoustic models in ASR. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2903–2907. ISCA.

Hiroki Ouchi and Yuta Tsuboi. 2016. Addressee and response selection for multi-party conversation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2133–2143, Austin, Texas. ACL.

Aleksei Pugachev, Oleg Akhtiamov, Alexey Karpov, and Wolfgang Minker. 2017. Deep learning for acoustic addressee detection in spoken dialogue systems. In *Conference on Artificial Intelligence and Natural Language (AINL)*, pages 45–53. Springer.

Björn Schuller, Stefan Steidl, Anton Batliner, Simone Hantke, Florian Hönig, Juan Rafael Orozco-Arroyave, Elmar Nöth, Yue Zhang, and Felix Weninger. 2015. The interspeech 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 478–482. ISCA.

Björn Schuller et al. 2017. The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3442–3446. ISCA.

Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Larry Heck. 2012. Learning when to listen: Detecting system-addressed speech in human-human-computer dialog. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 334–337. ISCA.

Elizabeth Shriberg, Andreas Stolcke, and Suman V Ravuri. 2013. Addressee detection for dialog systems using temporal and spectral dimensions of speaking style. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2559–2563. ISCA.

Ingo Siegert and Julia Krüger. 2018. How do we speak with Alexa - Subjective and objective assessments of changes in speaking style between HC and HH conversations. *Kognitive Systeme*, 1.

Ingo Siegert, Julia Krüger, Olga Egorow, Jannik Nietzold, Ralph Heinemann, and Alicia Lotz. 2018. Voice assistant conversation corpus (VACC): A multi-scenario dataset for addressee detection in human-computer-interaction using Amazon ALEXA. In *LREC 2018 Workshop "LB-ILR2018 and MMC2018 Joint Workshop"*, pages 51–54. ELRA.

Anastasiia Spirina, Olesia Vaskovskaia, Maxim Sidorov, and Alexander Schmitt. 2016. Interaction quality as a human-human task-oriented conversation performance. In *International Conference on Speech and Computer (SPECOM)*, pages 403–410. Springer.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204. IEEE.

TJ Tsai, Andreas Stolcke, and Malcolm Slaney. 2015. A study of multimodal addressee detection in human-human-computer interaction. *IEEE Transactions on Multimedia*, 17(9):1550–1561.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2017. Mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412.

Rui Zhang, Honglak Lee, Lazaros Polymenakos, and Dragomir Radev. 2018. Addressee and response selection in multi-party conversations with speaker interaction RNNs. In *AAAI Conference on Artificial Intelligence*, pages 5690–5697. AAAI.