# Discrimination of Linguistic and Non-Linguistic Vocalizations in Spontaneous Speech: Intra- and Inter-Corpus Perspectives

*Felix Weninger, Björn Schuller*

Institute for Human-Machine Communication, Technische Universität München, Germany

(weninger|schuller)@tum.de

## Abstract

We present a large-scale study on classification of linguistic and non-linguistic vocalizations including laughter, vocal noise, hesitation and consent on four corpora amounting to 46 h of spontaneous conversational speech. We consider training and testing on speaker-independent subsets of single corpora (intra-corpus) as well as inter-corpus experiments where models built on one or more corpora are evaluated on a disjoint corpus. Our results reveal that while inter-corpus performance is considerably lower than comparable intra-corpus results, this effect can be countered by data agglomeration; furthermore, we observe that inter-corpus classification accuracies indicate suitability of corpora for building generalizing models.

## 1. Introduction

Recognizing paralinguistic information, such as the emotion or intent, of the dialog partners is considered to be vital for human-machine interaction resembling natural human-human conversations. Notably, non-linguistic vocalizations such as laughter or sighs are one of the most important channels for conveying such paralinguistic information, as they can be decoded very robustly by humans [1]. Hence, it is crucial for affect-sensitive technical systems to recognize such vocalizations robustly, and to discriminate them from linguistic vocalization, i. e., spoken words. In this study, we present a generic, purely data-based approach for discrimination of five different classes of non-linguistic and linguistic vocalizations, in contrast to many studies focusing on detection of a single type of vocalization such as laughter [2]. Furthermore, we extend previous studies on data-based vocalization classification, such as [3–5], which are limited to evaluation within single corpora, i. e., *intra-corpus* evaluation, by adding an *inter-corpus* perspective. Such inter-corpus testing, taking into account different acoustic conditions or conversation domains, is highly relevant: When building audio recognition systems for real-life use, training data matching the specific application scenario might not be available. Consequently, inter-corpus testing has been widely addressed in the field of automatic speech recognition, e. g., in [6]. Besides, in the non-linguistic vocalization domain similar evaluation protocols have been followed in [7] for a binary classification task (to tell apart laughter from speech), indicating notable performance differences due to varying recording conditions.

Motivated by these first results, we now present a large-scale study on intra- and inter-corpus classification of linguistic vocalizations, including four corpora of spontaneous speech (cf. Section 2) for a total of 46 h of speech. Following previous inter-corpus studies in emotion recognition such as [8] we also address training data agglomeration for inter-corpus testing, to answer whether there really is 'no data like more data'. The consideration of large-scale, speaker-independent, inter-corpus evaluation on the one hand comes with some restrictions on the other hand: First, we limit this study to classification of pre-segmented data, as detection of non-linguistic events in naturalistic speech recordings can be very challenging even in homogeneous conditions [4, 7]; second, as different labeling schemes have to be unified given that some classes (e. g., coughing) are sparsely or not at all annotated in some corpora, we decided to define a rather coarse five-class task to discriminate segments corresponding to: words (linguistic vocalization); laughter; vocal noise including breathing, sighing or coughing; non-verbal consent ('mhm'); and filled pauses ('um', 'uh'). Classifier parameterization and experimental protocols are described in Section 3 before presenting and discussing results in Sections 4 and 5.

## 2. Databases

In the following we describe our evaluation databases of spontaneous speech. Their language is English, yet featuring various accents and dialects (see below).

### 2.1. AVEC subset of the SEMAINE database

As a first spontaneous speech database, we selected the official corpus of the 2011 Audio/Visual Emotion Challenge (AVEC) [9], which is part of the SEMAINE corpus [10] (http://semaine-db.eu). For the AVEC corpus conversations between humans and operators pretending to be emotionally intelligent virtual agents were recorded. Operators were instructed to create a conversation as natural as possible while playing emotionally stereotyped 'characters' and restraining to stock phrases keyed to the user's emotional state, simulating no full language understanding. Audio was recorded at 48 kHz with 24 bits per sample. Transcribed non-linguistic vocalizations include laughter, sighs and breathing. The partitioning used in this study conforms to the Challenge [9]: The training partition contains 31 recording sessions, while the development and test partitions contain 32 sessions. Only the user's speech is considered. All subjects speak English fluently, with about 3/4 of the subjects being native speakers, with a prevalence of Irish background.

### 2.2. TUM AVIC

Secondly, we used the "TUM AVIC" corpus [11] which has also been the basis for the Affect Sub-Challenge of the Interspeech 2010 Paralinguistic Challenge [12]. In the scenario setup, an experimenter plays the role of a product presenter and leads the subject through a commercial (car) presentation. The subject's role is to listen to and actively interact with the experimenter considering his/her interest in the addressed topics. We exclusively use speech data recorded by the lapel microphone

| corpus | len [h] | laugh | hesit. | cons. | v.noise | word | total | # subj. | accent/dialect | scenario |
|---|---|---|---|---|---|---|---|---|---|---|
| AVEC | 6.8 | 356 | 1 175 | 41 | 97 | 35 842 | 37 511 | 17 | Irish | human-agent conv. |
| AVIC | 2.3 | 294 | 1 204 | 344 | 1 290 | 16 441 | 19 573 | 21 | German | product present. |
| Buckeye | 26 | 1 874 | 5 594 | 939 | 20 504 | 231 422 | 260 333 | 40 | Columbus, Ohio | interview |
| COSINE | 11 | 3 267 | 984 | 374 | 1 313 | 70 585 | 76 523 | 37 | various U.S. | multi-party conv. |

Table 1: Four spontaneous (English) speech corpora: length (len) of considered recordings; number of instances in each of the 'laugh(ter)', 'hesit(ation)', 'cons(ent)', 'v(ocal) noise' and 'word' classes; number of subj(ects); prevalent accent / dialect of subjects; and recording scenario.

(44.1 kHz, 16 bit) as in the Challenge. 21 subjects took part in the recordings, three of them Asian, the remaining European. The language throughout experiments is English, and all subjects are non-native, yet very experienced English speakers, most of them with a German background. The mean age of the participants is 29.9 years. The total recording time is 10.4 h; in this study, we only use the subjects' turns (2.3 h). Non-linguistic vocalizations have been explicitly labeled in the transcriptions using markers for breathing, consent ('mhm'), hesitation ('um', 'uh'), laughter, and coughing as well as other human noise. As in [12], the speech data from the 21 speakers were split into a training, development, and test set in a speaker independent way trying to achieve the best possible balance with respect to gender, age, and ethnicity and following roughly a 40 / 30 / 30 % partitioning.

### 2.3. Buckeye

The third and largest spontaneous speech corpus considered in this study is the Buckeye corpus [13], which contains recordings of interviews with 40 subjects who are natives of Central Ohio. Interviews were conducted in a small seminar room. The speech of the subjects was recorded with a head mounted microphone while the interviewer did not wear a microphone. Thus, only the subject's speech is intellegible in the recordings, and we only use that data in this study (roughly 26 h out of 38 h). It was found that the formality of the interview dissipated quickly into a friendly conversation [13]; thus, the speech is highly spontaneous and contains a variety of non-linguistic vocalizations. Locations of laughter and vocal noise are marked, and filled pauses and backchannels are transcribed phonetically. As the authors of the corpus do not prescribe an experimental setup, we follow the partitioning from our previous study [4] on non-linguistic vocalizations to divide the corpus into a training, development, and test set, stratified by age and gender.

### 2.4. COSINE

Finally, the rather novel COSINE corpus [14] was taken into account as an example of multi-party conversations recorded in real world environments. The recordings were captured on a wearable recording system so that the speakers were able to walk around during recording. Since the participants were asked to speak about anything they liked and to walk to various noisy locations, the corpus consists of natural, spontaneous, and highly disfluent speaking styles partly masked by indoor and outdoor noise sources such as crowds, vehicles, and wind. The recordings were captured using multiple microphones simultaneously, however, to match most application scenarios, we exclusively used speech recorded by a close-talking microphone (Sennheiser ME-3). We used all ten transcribed sessions, containing 11.40 h of pairwise conversations and group discussions. All 37 speakers are fluent, but not necessarily native English speakers covering a broad range of United States accents. Each speaker participated in only one session and the speakers' ages range from 18 to 71 years (median 21 years). Laughter and vocal noise segments are marked in the transcription. For our experiments, we split the corpus according to the recommendation of the corpus' authors into a test set (sessions 3 and 10, 1.81 h of speech) and training set (remaining eight sessions).

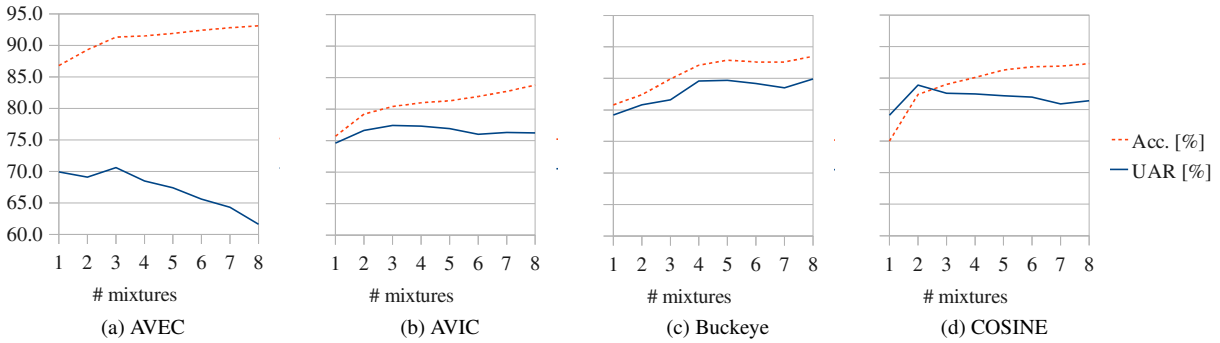
## 3. Experimental Setup

### 3.1. Preprocessing and Class Definition

To subdivide the corpora in word-like units for analysis, we used forced alignment segmentations of the databases by triphone Hidden Markov Models (HMMs) except for the Buckeye corpus which is delivered with a more advanced, yet fully automatic word-level segmentation [13]. Then, in order to perform inter-corpus experiments, corpus-specific mappings of word-like units in the transcriptions to the five classes words (linguistic vocalization), non-verbal backchannels indicating consent, filled pauses (hesitation), vocal noise, and laughter, were defined. In particular, for the AVIC database we unified the coughing, other human noise, and breathing classes to the vocal noise class; in the AVEC database, breath and sigh were unified to the vocal noise class. In the three corpora (all except AVIC) where consent and hesitation are transcribed phonetically without using special markers, we considered the segments labeled as 'mhm(mm)', 'aha', 'um-hum', 'um-hmm', 'mm-hmm' or 'aha' as consent, and those labeled as either 'em', 'eh(h)', 'um' or 'uh' as hesitation. While segments where laughter and vocal noise coincide with speech are annotated explicitly in Buckeye and COSINE, we considered them as 'words' for the purpose of this study; first, for consistency with the two other corpora where this phenomenon is not taken into account in the transcription; second, since such segments contain linguistic information which might be decoded by a speech recognizer in a second step. We removed all segments with a length below 100 ms, as such short segments often indicate alignment errors. The resulting number of instances per class in each of the four corpora is shown in Table 1. As expected, classes are heavily unbalanced; furthermore, it is notable that the 'consent' class is highly underrepresented in the AVEC corpus, probably due to the recording scenario designed to elicit strongly emotional responses from the user instead of simple backchanneling.

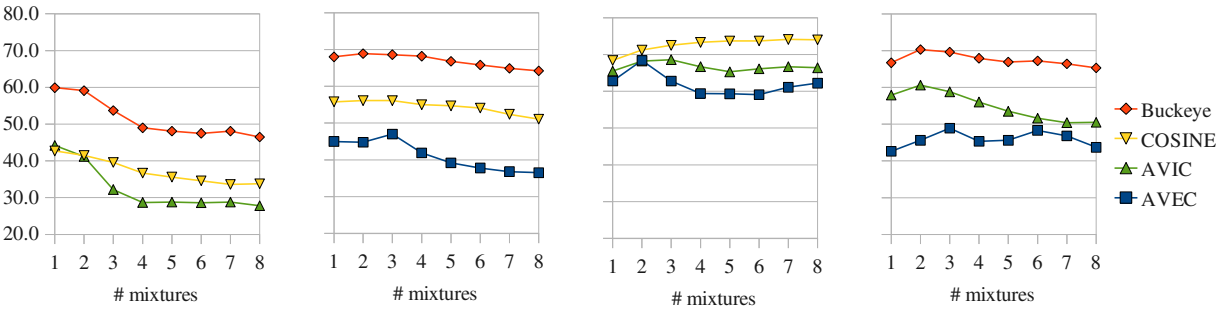
### 3.2. Classifier Setup

We considered isolated HMMs exclusively in this study, assigning the class corresponding to the model with the maximum likelihood (ML): Since there are no transitions between models, the a-priori class probabilities do not affect the decision, accounting for the class imbalance. We chose a strictly linear left-right topology with eight emitting states. HMMs were trained by six initial Expectation-Maximization (EM) iterations, after which additional Gaussian mixtures were added consecu-




(a) AVEC (b) AVIC (c) Buckeye (d) COSINE

Figure 1: Intra-corpus classification: Accuracy (Acc.) and unweighted average recall (UAR) in speaker-independent training and testing on AVEC (a), AVIC (b), Buckeye (c) or COSINE (d). Train / test partitioning according to Section 2.




(a) train on AVEC (b) train on AVIC (c) train on Buckeye (d) train on COSINE

Figure 2: Pair-wise inter-corpus classification: UAR [%] in training on either AVEC (a), AVIC (b), Buckeye (c) or COSINE (d), and testing on each of the three remaining corpora.

tively and re-estimated during four EM iterations, until the final models had eight Gaussian mixtures per state. We measured classification accuracy for each number of mixtures separately, as shown below. As features, we extracted the Mel frequency cepstral coefficients (MFCCs) 1–12 along with energy and their first and second order deltas from 26 Mel filter banks spanning 20–8 000 Hz. Notably, following [3], we did not employ cepstral mean normalization as this technique seems to require longer units of analysis, such as speaker turns, to perform robustly.

### 3.3. Intra- and Inter-Corpus Classification

We compared performance of intra-corpus classification with inter-corpus classification. For the former, we trained models on the defined training and development sets of each of the four corpora, then evaluated on the test set, thereby strictly enforcing speaker independence—in case of AVEC, we trained on the training and evaluated on the development set since non-linguistic vocalizations are extremely scarce in the test set. For the latter, two scenarios were considered: Pairwise inter-corpus classification, i. e., selecting one (complete) corpus for training and another for testing, and a leave-one-corpus-out (LOCO) strategy where for each testing corpus the remaining three corpora were jointly used as training material.

## 4. Results

Our experiments are evaluated in terms of accuracy and unweighted average recall (UAR) of the five classes; the first corresponds to a system's rate of correct classifications while the second reflects its ability to discriminate the classes.

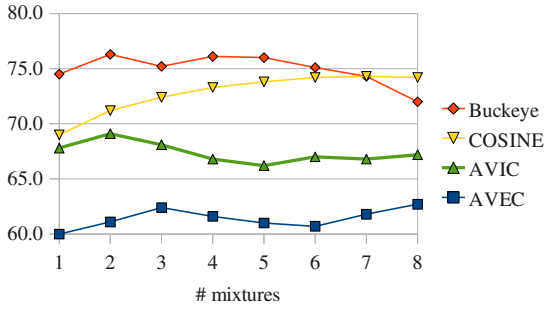



Figure 3: Leave-one-corpus-out (LOCO) inter-corpus classification: UAR [%] in training on the union of three corpora and testing on the remaining one.

### 4.1. Intra-Corpus Classification

As an indicator of how challenging the proposed five-class task is on each of the four speech corpora, we show the performance of intra-corpus classification in Figures 1a through 1d. It can be seen that an increased number of Gaussian mixtures per state constantly increases accuracy—mainly due to an increased recall of the majority 'word' class, but at the expense of lower recall of the minority classes—for instance, recall of consent drops down to zero for the AVIC corpus at 8 mixtures. Thus, accuracy stands in contrast to UAR, which appears to be optimal at a low ($\leq$ 3) number of mixtures except for the Buckeye corpus (where more training data for the 'smaller' classes is available). The latter is also the corpus with highest overall UAR (82.9 % on average over 1–8 mixtures). At the other end of the scale we find the AVEC corpus (67.1 % average UAR), which can be attributed to data sparsity.

| Test on | pair-wise | LOCO |
|---|---|---|
| AVEC | 49.6 | 61.4 |
| AVIC | 51.4 | 67.4 |
| Buckeye | 62.0 | 74.9 |
| COSINE | 54.8 | 72.8 |

Table 2: Pair-wise and LOCO inter-corpus classification: Expected UAR [%] for four different testing corpora.

### 4.2. Inter-Corpus Classification

This picture of 'difficulty' of the corpora does not change when adding the inter-corpus perspective: Again, it is clearly visible that recognition on Buckeye performs best, followed by COSINE, AVIC and finally AVEC—this ranking is consistent throughout all iterations of our pair-wise inter-corpus experiments, involving four different training corpora and 1–8 Gaussian mixtures per HMM state, as shown in Figures 2a through 2d. Thus, apparently the corpora that can be classified most robustly also deliver the best generalizing models as indicated by inter-corpus testing. In particular, in Figures 2a, 2b and 2d we have evidence that it is not the sheer size of the Buckeye corpus that makes it suitable for building robust models (2c), but rather its prototypicality—indicated by high classification performance on Buckeye when training with other corpora.

Furthermore, the intra-corpus accuracies obtained on the corpora are clearly correlated with the ability to build generalizing models from them; this effect cannot be simply attributed to the different sizes of the databases, since similar results are obtained by training on either AVIC or COSINE (Figures 2b, 2d); in contrast, this could be due to effects of recording noise in COSINE. Finally, although results are not directly comparable, the general picture is that the inter-corpus classification performance is considerably lower than the intra-corpus one, with a result of 74.1 % UAR (training on Buckeye, testing on COSINE, 7 mixtures) as an upper bound. Notably, there seems to be a large 'incompatibility' between AVIC and AVEC, as the UAR drops below 30 % when training on AVEC and testing on AVIC.

On the contrary, evaluation in a LOCO fashion (Figure 3) yields much more stable results, avoiding the above-mentioned performance drop, with a lower bound of 60.0 % UAR (testing on AVEC with 1 mixture). Overall, this kind of evaluation corroborates the results as to which corpora are 'easiest' to classify, but also demonstrates the benefit of data agglomeration.

The latter is corroborated by the statistical perspective on inter-corpus classification which is given in Table 2: In reality, it is unknown which training data and which classifier parameterization (here, the number of mixtures per HMM state) are optimal for some unlabeled test data. Thus, we compare the expected UAR on each of the four corpora considered, when selecting an arbitrary number of mixtures between 1 and 8, and (a) an arbitrary (disjoint) corpus—corresponding to pair-wise inter-corpus evaluation—or (b) the union of the three remaining corpora, cf. LOCO evaluation. From Table 2, it can be seen that agglomeration of training corpora as done in LOCO evaluation delivers significant gains over single-corpus training for all of the test corpora considered ($p < 0.001$ according to a z-test).

## 5. Conclusions

In a large-scale evaluation on four spontaneous speech corpora, we have demonstrated that inter-corpus discrimination of non-linguistic vocalizations in spontaneous speech is a challenging task: Generally, when using single corpora for training, one would expect a drastic decrease in performance compared to intra-corpus evaluation. It is promising, though, that through data agglomeration, this phenomenon can be mitigated to some extent, and that classification performance on a corpus seems to be a proxy for its ability to build generalizing models, which is hard to predict in practice. Interesting directions for future research can be found in semi-supervised learning: In a first step, intra-corpus data could be included via unsupervised adaptation of the agglomerated data sets; next, co-training with unlabeled spontaneous speech from the internet could be investigated. Finally, data agglomeration to alleviate cross-cultural effects in vocalization classification should be considered.

## 6. References


[1] S. T. Hawk, G. A. van Kleef, A. H. Fischer, and J. van der Schalk, "'Worth a thousand words': absolute and relative decoding of nonlinguistic affect vocalizations," *Emotion*, vol. 209, no. 3, pp. 293–305, 2009.

[2] M. Knox and M. Mirghafori, "Automatic laughter detection using neural networks," in *Proc. of INTERSPEECH*, Antwerp, Belgium, 2007, pp. 2973–2976.

[3] B. Schuller, F. Eyben, and G. Rigoll, "Static and Dynamic Modelling for the Recognition of Non-Verbal Vocalisations in Conversational Speech," in *Perception in Multimodal Dialogue Systems, Proc. of PIT 2008*. Springer, 2008, pp. 99–110.

[4] F. Weninger, B. Schuller, M. Wöllmer, and G. Rigoll, "Localization of non-linguistic events in spontaneous speech by non-negative matrix factorization and long short-term memory," in *Proc. of ICASSP*, Prague, Czech Republic, 2011, pp. 5840–5843.

[5] N. Campbell, J. Kane, and H. Moniz, "Processing 'yup!' and other short utterances in interactive speech," in *Proc. of ICASSP*, Prague, Czech Republic, 2011, pp. 5832–5835.

[6] S. Tskakalidis and W. Byrne, "Acoustic training from heterogeneous data sources: Experiments in mandarin conversational telephone speech transcription," in *Proc. of ICASSP*, Philadelphia, USA, 2005, pp. 461–464.

[7] S. Petridis, A. Asghar, and M. Pantic, "Classifying laughter and speech using audio-visual feature prediction," in *Proc. of ICASSP*, Dallas, USA, 2010, pp. 5254–5257.

[8] I. Lefter, L. J. M. Rothkrantz, P. Wiggers, and D. A. van Leeuwen, "Emotion recognition from speech by combining databases and fusion of classifiers," in *Proc. of Text, Speech and Dialogue*, Berlin, Germany, 2010, pp. 353–360.

[9] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011—The First International Audio/Visual Emotion Challenge," in *Proc. First International Audio/Visual Emotion Challenge and Workshop (AVEC) held in conjunction with ACII*, ser. LNCS. Memphis, TN, USA: Springer, 2011, pp. 415–424.

[10] G. McKeown, M. Valstar, M. Pantic, and R. Cowie, "The SEMAINE corpus of emotionally coloured character interactions," in *Proc. of ICME*, Singapore, 2010, pp. 1079–1084.

[11] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application," *Image and Vision Computing Journal*, vol. 27, no. 12, pp. 1760–1774, 2009.

[12] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge – Age, Gender, and Affect," in *Proc. of Interspeech*, Makuhari, Japan, 2010, pp. 2794–2797.

[13] M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, *Buckeye Corpus of Conversational Speech (2nd release)*. Columbus, OH, USA: Department of Psychology, Ohio State University (Distributor), 2007, [www.buckeyecorpus.osu.edu].

[14] A. Stupakov, E. Hanusa, D. Vijaywargi, D. Fox, and J. Bilmes, "The design and collection of COSINE, a multi-microphone in situ speech corpus recorded in noisy environments," *Computer Speech and Language*, vol. 26, no. 1, pp. 52–66, 2011.