Static and Dynamic Modelling for the Recognition of Non-verbal Vocalisations in Conversational Speech

Björn Schuller, Florian Eyben, and Gerhard Rigoll

Institute for Human-Machine Communication, Technische Universität München,
Theresienstrasse 90, 80333 München, Germany
{sch, eyb, ri}@mmk.ei.tum.de
http://www.mmk.ei.tum.de

Abstract. Non-verbal vocalisations such as laughter, breathing, hesitation, and consent play an important role in the recognition and understanding of human conversational speech and spontaneous affect. In this contribution we discuss two different strategies for robust discrimination of such events: dynamic modelling by a broad selection of diverse acoustic Low-Level-Descriptors vs. static modelling by projection of these via statistical functionals onto a 0.6k feature space with subsequent de-correlation. As classifiers we employ Hidden Markov Models, Conditional Random Fields, and Support Vector Machines, respectively. For discussion of extensive parameter optimisation test-runs with respect to features and model topology, 2.9k non-verbals are extracted from the spontaneous Audio-Visual Interest Corpus. 80.7% accuracy can be reported with, and 92.6% without a garbage model for the discrimination of the named classes.

1 Introduction

Speech is an essential part of human to human communication. It is perhaps the most natural way for people to exchange information with each other. Therefore, if we want machines that are able to communicate with us via natural speech communication, we need robust and intelligent methods for speech recognition, speech understanding and speech synthesis.

Speech recognition research in the past has mainly focused on well defined recognition tasks. These tasks had a restricted vocabulary and task specific stochastic or rule-based grammar. The utterances used for evaluation were mostly read by native speakers under perfect acoustic conditions. Main focus was laid on phoneme based word recognition. Near perfect recognition results have been reported for such tasks (under laboratory conditions) already more than a decade ago [Young, 1996].

Such a system will, however, not work well for spontaneous, conversational speech in applications like dialog systems, call centre loops or automatic transcription systems for meetings. This is due to various non-verbal sounds and irregularities encountered in spontaneous speech. These include disfluencies (filled

E. André et al. (Eds.): PIT 2008, LNCS 5078, pp. 99–110, 2008.

and unfilled pauses, corrections and incomplete words), interjections (e.g. laughing, crying, agreement/disagreement: "aha/ah ah"), human noises (e.g. yawning, throat clearing, breathing, smacking, coughing, sneezing) and other sounds like background conversation or noise [Ward, 1991].

Read speech conveys only the information contained in the spoken words and sentences. In contrast, spontaneous speech contains more extralinguistic information "between the lines", including irony, speaker emotion, speaker confidence and interest in conversation [Schuller et al., 2007]. Next to prosodic features [Kompe, 1997], disfluencies and non-verbal clues such as filled pauses, laughter or breathing reveal much about this extralinguistic information [Schuller et al., 2007]. An automatic spontaneous speech recognition system will only be able to detect information carried on the verbal level. For understanding the extralinguistic information carried by spontaneous speech, non-verbal information is vital [Campbell, 2007], [Decaire, 2000], [Lickley et al., 1991].

A speech recogniser that is able to understand the meaning of spoken language to some extent, must be capable of spotting non-verbal sounds and identifying their type. In contrast to some previous work, which aimes at detection of non-verbal sounds in order to improve robustness of speech recognition [Schultz and Rogina, 1995], this paper deals with the explicit identification of the type of non-verbal vocalisation. [Schultz and Rogina, 1995] only reports on increase in Word Accuracy, and not on correct identification of non-verbal sounds.

The article is structured as follows: in section 2 existing work is discussed, in section 3 details on the database are provided, in section 4 the proposed methods are introduced before results and conclusions in section 5 and section 6, respectively.

2 Existing Work

Various work exists on automatic recognition of few types of Non-Verbals. Covered are especially filled pauses and laughter [Kennedy and Ellis, 2004], [Truong and van Leeuwen, 2005].

Filled pauses. A filled pause detection system was introduced by M. Goto et al. in [Goto et al., 1999]. A quick summary of the technique is given in the following: the system is able to spot two hesitation phenomena, namely filled pauses and word lengthening, in a continuous stream of spontaneous speech. The basic assumption is that hesitations are uttered when the "speaking process is waiting for the next speech content from the thinking process" and thus the speaker cannot change articulatory parameters in that instant. A voiced sound with nearly constant fundamental frequency (F_0) and minimal spectral envelope deformation over time will be produced. The system detects such voiced sounds with minimal variation in the articulatory parameters, which it assumes to be hesitations. A recall rate of 84.9% percent and a precision rate of 91.5% is reported for the spotting of hesitations.

Laughter. For laughter detection and especially synthesis of laughter various papers have been published by N. Campbell et al. [Campbell et al., 2005]. The basic approach they take is the following: an excessive study about various types of laughter has been conducted. It has been found that laughter consists of four distinguishable basic segments, namely voiced laugh, chuckle, ingressive breathy laugh, and nasal grunt. Hidden-Markov Models (HMM) are trained on these laughter segments. A language model, defining in what sequence laughter segments occur, is further given. A success rate of 81% compared to hand labelled data is obtained in detecting the correct laughter segments. 75% success rate is reported when using the grammar to detect laughter type based on the detected laughter segments. This approach is very suitable for detecting laughter types, once a speech/ laughter discrimination has been performed. The latter, however, is not described in that paper.

Another, quite recent, approach [Knox and Mirghafori, 2007] is presented by M. Knox. Experiments are carried out on a large English Meeting room database (ICSI Meetings database). Features from 25ms frames with a forward shift of 10ms are extracted. 75 consecutive frames are fed as input to a neural network, which assigns a target (speech/laughter) to the centre frame. The output of several such neural networks, operating on different feature sets is again fed into a combiner neural network to produce the final output. In this way a target is assigned to every frame. An Equal-Error Rate of 7.9% is achieved. The paper investigates several feature sets whereby Mel-Frequency Cepstral Coefficient (MFCC) based ones give the best results.

The detection of other sounds, such as breathing, yawning or throat clicking is yet quite unexplored. Further, no work is known to us that approaches the problem in a strictly data-driven manner, independent of the type of non-verbal vocalisation to be detected.

In this paper we will therefore focus on the data-driven detection of non-verbal sounds in general. Various dynamic and static classification methods for discriminating between different classes of isolated Non-Verbals are discussed and evaluated.

3 Database

In our experiments we use a database containing 2.9k isolated Non-Verbals extracted from the Audio-Visual Interest Corpus (AVIC) [Schuller et al., 2007]. The AVIC database contains human conversational speech of 21 subjects (10 of them female, 3 of them Asian, others European with balanced age classes) discussing in English with a product presenter who leads them through a commercial presentation. Voice data is recorded by a headset, and a condenser far-field mic (approx. 50cm distance) at an audio sampling rate of 44.1kHz with 16Bit quantisation. All presenter comments are not included in the extracted segments because this would perturb the balanced distribution of number of utterances among speakers. Thus, the total recording time for males resembles 5:14:30h with 1,907 turns, for females 5:08:00h with 1,994 turns, respectively. The lengths of

the utterances range from 0.1s to 10.9s with 2.1s on average. Apart from five levels of interest, the spoken content including non-verbal interjections on a word level is transcribed. These interjections are *breathing*, *consent*, *coughing*, *hesitation*, *laughter*, *long pause*, *short pause*, and *other human noise*. Tab. 1 shows the distribution among classes of the Non-Verbals used in the ongoing, whereby coughing was excluded due to sparse instances. Other human noise was mapped onto the new class *garbage* for the following experiments. Further, of the 4.503 extracted Non-Verbals by Forced Alignment, only such having a minimum of 100 ms are kept. Non-Verbals shorter that 100ms have in most cases been incorrectly aligned. Moreover, feature extraction and model evaluation on such short segments is very error prone. The maximum length of occurring Non-Verbals is 2s. Each turn contains between 0 and 31 words with 4.71 words on average (silence and Non-Verbals are hereby not counted). Of the 3,901 turns in total, 2,710 contain between 1 and 7 Non-Verbals. Likewise, there is a total of 18,581 spoken words, and 23,084 word-like units including Non-Verbals (19.5%).

Table 1. Distribution of the Non-Verbals in *AVIC* across the 5 classes

Breathing	**Consent**	**Garbage**	**Hesitation**	**Laughter**	**TOTAL**
452	325	716	1,147	261	**2,901**

4 Proposed Method

In this section we investigate three different methods for the discrimination between 5 classes of Non-Verbals, namely Breathing, Consent, Garbage, Hesitation, and Laughter: Hidden Markov Models (HMM), Hidden Conditional Random Fields (HCRF), and Support Vector Machines (SVM).

Extensive tests are conducted for HMM in order to find an optimal configuration (features and model topology) for the task at hand. These are described in more detail in Sect. 4.1. The HCRF are initialised with the parameters of corresponding trained HMM, and thus are fully comparable to the HMM [Gunawardana et al., 2005]. Six feature sets, based on MFCC and PLP are evaluated in conjunction with the two dynamic classifiers, HMM and HCRF. For static classification with SVM a large feature set based on acoustic low-level descriptors (LLD) is used, which has successfully been used in the field of paralinguistics [Schuller et al., 2008]. The following sections describe each of the three methods in more detail.

4.1 Non-verbals Recognition Using HMM

No previous evaluations for the task of Non-Verbals discrimination regarding HMM topology optimisation have been conducted. Therefore, we must find an optimal topology for the task. In phoneme based speech recognisers HMM with

Table 2. Description of the six feature sets for dynamic classifiers (HMM and HCRF)

Set	Features	Dimension
$MFCC_E$	12 MFCC (1-12) + E + δ + $\delta\delta$	39
$MFCC_0$	13 MFCC (0-12) + δ + $\delta\delta$	39
$MFCC_0^{cms}$	13 MFCC (0-12) + δ + $\delta\delta$ (after Cepstral Mean Subtraction)	39
$PLPCC_E$	12 PLPCC (1-12) + E + δ + $\delta\delta$	39
$PLPCC_0$	13 PLPCC (0-12) + δ + $\delta\delta$	39
$PLPCC_0^{cms}$	13 PLPCC (0-12) + δ + $\delta\delta$ (after Cepstral Mean Subtraction)	39

3 states are used to model phonemes. Non-Verbals can be longer than phonemes (see Sect. 3) and have more acoustic variation. It can thus be assumed that more than 3 states are required for Non-Verbals HMM.

One can approach the task of HMM topology optimisation in many ways. An optimal solution will however only be found if all possible combinations of topology parameters for all classes are evaluated. The topology parameters of interest are: number of emitting states (N), number of Gaussian mixture components (M) for each state's output distribution, and the state transition configuration ($\mathbf{A}$), i.e. which transitions between which states are allowed. Due to the exponentially large amount of evaluations required for finding an optimal topology, such exhaustive search is not computable. In order to get an idea of how the HMM topology and choice of features affects the results, a small set of parameters will therefore be tested. The results can be used in future work to further optimise the model topology. The detailed evaluation procedure, including feature and parameter sets, is described in the following three subsections. Results are given in Sect. 5.

Feature sets. Six feature sets based on Mel-Frequency Cepstral Coefficients (MFCC), and Perceptual Linear Predictive Cepstral Coefficients (PLPCC) [Hermansky, 1990] are evaluated. An overview is given in Tab. 2. All features are extracted from frames of 25ms length sampled at a rate of 10ms. A Hamming window is applied to the frames before transformation to the spectral domain. Using a Mel filter bank of 26 channels, 13 MFCC, and 13 PLPCC including the 0^{th} coefficient are computed. Also, the log-energy E is computed for every frame. First and second order regression coefficients are appended to all six feature sets. Cepstral Mean Subtraction (CMS) is applied to one MFCC and one PLP based feature set. This means, for each cepstral coefficient the mean is computed over all corresponding coefficients in the input and then subtracted.

HMM topology parameters. Three different types of HMM structure are investigated: the first being a linear HMM, i. e. a left-right HMM with no skip-state transitions (only transitions from state n to states n and $n+1$ are allowed). The second being a Bakis topology HMM (left-right) with one skip-state transition ($N_{skip} = 1$), i. e. with allowed transitions from a state n to states n, $n + 1$

and $n+2$. The number N of emitting states is varied from 1 to 10. Each number of states is tested with $M = 1$ and $M = 8$ Gaussian mixture components. This results in a total of 40 different sets of model parameters to be evaluated.

Evaluation procedure. The 40 parameter sets introduced in the previous section are evaluated for all six feature sets independently. Thus, a total of 240 evaluations is conducted. Each single evaluation is performed in a speaker independent 3-fold stratified cross-validation (SCV). For the SCV, the AVIC data set is split into three speaker disjunctive parts, each containing data from one third of the speakers. Parts 1 through 3 are used for testing in folds 1 through 3, respectively. The remaining two parts are used for training. Splitting by speakers, however, introduces some problems that one must be aware of: the types of Non-Verbals and the number of Non-Verbals are not equally distributed among speakers (see Sect. 3 for more details). For example, some speakers are more fluent or confident and thus produce fewer hesitations. Therefore, among the folds and the classes in each fold there will be notable differences in the amount of test data vs. training data. Tab. 3 shows the number of training and test instances for each fold. HMM with $M = 1$ are trained in 4 iterations of

Table 3. Number of occurrences of each class in test and training data for each fold

Fold	**[#] test**					
	Breathing	Consent	Garbage	Hesitation	Laughter	**Total**
1	129	95	126	340	68	**758**
2	83	88	264	281	100	**816**
3	240	142	326	526	93	**1327**
Fold	**[#] train**					
	Breathing	Consent	Garbage	Hesitation	Laughter	**Total**
1	323	230	590	807	193	**2143**
2	369	237	452	866	161	**2085**
3	212	183	390	621	168	**1574**

Baum-Welch re-estimation [Young et al., 2006]. Models with $M = 8$ are created from the trained models with $M = 1$ by successive mixture splitting, i. e. the number of mixture components M is doubled three times. After each doubling of M, 4 re-estimation iterations are performed. One model is trained for each Non-Verbal. The most likely model is found by Viterbi evaluation. Priors are integrated by the number of occurrences (in the training set) of the corresponding class.

A discussion of the results and best topology and feature set combination is given in Sect. 5.

4.2 Isolated Recognition of Non-verbals Using HCRF

Hidden Conditional Random Fields (HCRF) have become popular in the last couple of years. They have been successfully applied to tasks such as phone classification [Gunawardana et al., 2005] and meeting segmentation [Reiter et al., 2007]. They are an extension of the Conditional Random Fields first introduced by Lafferty et al. [Lafferty et al., 2001].

In this work we use HCRF that are initialised with the parameters of trained HMM. The HCRF are not trained any further in order to have a direct comparison of the two model types. Configurations with $N = 1..10$ hidden states and 1 and 8 Gaussian mixture components are examined. Only the feature set $PLPCC_E$ is investigated thoroughly because the best HMM recognition results are reported with this feature set. The same speaker independent 3-fold stratified cross-validation procedure as for HMM is used. The results for classification of Non-Verbals with HCRF are also given in Sect. 5.

4.3 Isolated Recognition of Non-verbals Using SVM

In this section a completely different approach for discrimination of different types of Non-Verbals is presented. The previous approach is based on dynamic models (HMM and HCRF) used in speech recognition applications because such models can be easily integrated into existing speech recognisers. Yet, alone for the task of distinguishing the type reliably in a second pass, after segmenting data into speech and Non-Verbal segments, for example, a static classification approach can be used.

Features extracted for the dynamic classifiers are sequences of feature vectors $\mathbf{x}_t$ with a sequence length proportional to the length of the input data. For static classification only one feature vector $\mathbf{x}$ with 622 features is extracted for each Non-Verbal utterance. The number of features is reduced to $D' = 108$ by a sequential forward floating search correlation-based (CFS) feature selection step. For computation of the 622 dimensional feature vector the low-level descriptors (LLD) given in Tab. 4 form the basis. Functionals are applied to the evolution of these LLD over time to obtain time and length independent static features. Statistical characteristics of LLD such as mean, median, minimum and

Table 4. Acoustic LLD used in computation of static feature vector

Type	**LLD**
Time Signal	Elongation, Centroid, Zero-Crossing Rate
Energy	Log-Frame-Energy
Spectral	0-250 Hz, 0-650 Hz, Flux Roll-Off + δ, Centroid + δ
Pitch	F_0 (fundamental frequency)
Formants	F1-F7 Frequency + δ, BW. + δ
Cepstral	MFCC 1-15 + δ + $\delta\delta$
Voice Quality	Harmonics to Noise Ratio (HNR)

maximum position and value, and standard deviation are used as functionals. These are also computed from first (δ) and second order ($\delta\delta$) regression coefficients of LLD, to better model LLD change over time. For more information see [Schuller et al., 2008].

The exact same speaker independent 3 fold partitioning for test and training as is used for dynamic classifiers in Sect. 4.1 is applied. Basing on previous experience in [Schuller et al., 2007], Support-Vector-Machines (SVM) trained with a Sequential Minimal Optimisation (SMO) algorithm are used as the classifier of choice.

5 Results

Table 5 compares the best results of each feature set. Overall best results are achieved with the feature set $PLPCC_E$. 80.7% of all instances are classified correctly when using linear topology HMM with $N = 9$ emitting states, and $M = 8$ Gaussian mixture components. 4 iterations are used for training models with $M = 1$. 12 additional iterations are required for models with $M = 8$.

Table 5. Discrimination between 5 Non-Verbals classes. HMM as classifier. 3-fold SCV, speaker independent. Best results (accuracies) and model parameters associated with best result for each feature set. Optimal topology is linear, each, meaning zero skip-states.

Parameters	$MFCC_0$	$MFCC_0^{cmn}$	$MFCC_E$	$PLPCC_0$	$PLPCC_0^{cmn}$	$PLPCC_E$
Best acc. [%]	79.5	74.2	77.7	79.3	73.4	79.7 (**80.7**[1])
N	8	7	5	10	9	**9**
M	8	8	8	8	8	**8**

PLP based features seem to require more states in the models for good results, MFCC based features give good results with 8 states. Overall, both feature kinds lead to similar classification results, so that they may be interchanged, in whatever way it is required by their application.

Cepstral Mean Normalisation (CMN) has not proven well throughout all experiments for isolated recognition of Non-Verbals. One explanation for this phenomenon might be as follows: isolated Non-Verbals are very short (< 2 s). Computing the mean over the cepstral feature vectors of short segments, which only contain one uttered sound (such as breathing), more likely leads to a biased mean, i. e. not the long term mean related to noise or recording location properties. Subtracting this biased mean leads to loss of information and thus to lower recognition accuracies.

[1] This result was obtained by increasing M in steps of 1 instead of doubling M in each round of mixture increase.

Overall best results are obtained with the feature set $PLPCC_E$ using models with 8 Gaussian mixture components and linear topology. The Bakis topology has not proven well. Tab. 6 gives details on the results. Up to now, only 1 and 8 Gaussian mixture components were evaluated. The models with 8 mixture components are created from the trained models with 1 mixture component by doubling the number mixture components and applying 4 rounds of re-estimation after each increase of the number of mixture components. Likewise, only mixture component numbers M that are a power of 2 can be investigated. To evaluate the effect of M in more detail, configurations with $M = 1 - 16$ are analysed on the winning feature set $PLPCC_E$ using the winning linear HMM topology with $N = 9$. The number of mixture components is now increased in steps of 1. After each step 4 rounds of re-estimation are applied. Fig. 1 visualises the results. The

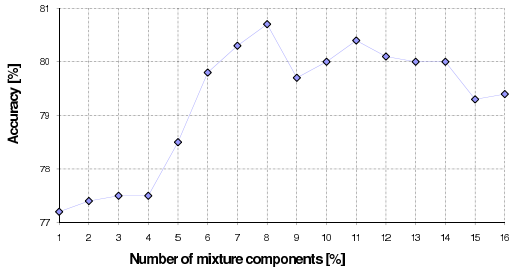

Fig. 1. Discrimination between 5 Non-Verbals classes. HMM as classifier. 3-fold SCV, speaker independent. Feature set $PLPCC_E$. $N = 9$, linear topology. Accuracies obtained with $M = 1 - 16$.

best result is obtained with $M = 8$ is **80.7%** accuracy. With HMM with the same configuration only 79.7% are obtained when creating models with $M = 8$ in fewer steps (i. e. by doubling the number of mixture components instead of increasing it by 1). This shows that more re-estimation iterations during mixture increasing are beneficial and lead to more accurate models.

Table 6. Discrimination between 5 Non-Verbals classes. HMM as classifier. 3-fold SCV, speaker independent. Feature set $PLPCC_E$. Selected numbers of states N vs. number of mixtures M and topology type (linear/Bakis).

[%] correct	$\boldsymbol{M = 1}$		$\boldsymbol{M = 8}$	
	linear	**Bakis**	**linear**	**Bakis**
$N = \mathbf{1}$	68.4	68.4	73.7	73.7
$N = \mathbf{3}$	73.5	72.3	75.5	74.8
$N = \mathbf{5}$	76.3	74.4	77.2	75.2
$N = \mathbf{7}$	77.0	74.8	77.6	76.1
$N = \mathbf{9}$	77.2	73.9	**79.7**	77.8

In order to better understand the sources of classification errors, we now take a look at the confusion matrix. A larger number of confusion matrices has been produced during the evaluations, however, all show one clear tendency, which can be seen in the exemplary confusion matrix shown in table 7: most confusions are related to the garbage model. This is most likely caused by the unspecific nature of the data in the garbage class and poor labeling. The class includes background noises (which, for example, have similar spectral characteristics than breathing), background speech, or speech segments which did not correspond to any word or Non-Verbal. The latter could be most likely confused with hesitation, which in practice would not be too wrong. To better model the individual classes of noise, more than one garbage model and more training data for each of these classes is required. To assess what the performance could be, if a better garbage modelling were available (i. e. more specific annotations and/or separate classes for the individual types of sounds in the garbage class), a test run without the garbage class has been conducted on the winning feature set $PLPCC_E$. As expected, results improve by 5% to 10% in this test run. Using models with 1 mixture component, 89.5% of the Non-Verbals are classified correctly. With 8 mixture components, **92.6%** are classified correctly.

Table 7. Confusion matrix: dynamic discrimination between 5 Non-Verbals classes. 3-fold SCV, speaker independent. Sum over all 3 folds. Optimal configuration: feature set $PLPCC_E$. $N = 9$, $M = 8$, linear topology. Mixture increase in steps of 1.

[#] classif. as →	**garbage**	**hesitation**	**consent**	**laughter**	**breathing**
garbage	**515**	93	22	41	41
hesitation	190	**929**	14	13	1
consent	28	37	**255**	3	3
laughter	17	1	2	**229**	12
breathing	18	2	1	19	**412**

Unlike the results obtained in [Reiter et al., 2007], the HCRF did not prove better than HMM for classification of Non-Verbals. The best result for HCRF - **77.8%** - is obtained with the configuration that gives best results for HMM: 9 states, 8 mixture components and linear topology.

With SVM **78.3%** of the instances are classified correctly in a 3-fold SCV. However, this again is below the best result, which is achieved with HMM. Consistent with the results for HMM, the confusion matrix for SVM reveals the garbage class as cause for most confusions. If these instances are ignored, the remaining 4 classes can be discriminated with an accuracy of **91.3%**.

6 Conclusion

We presented the robust recognition of 5 types of Non-Verbals, herein. Diverse models and feature-types were outlined and extensively evaluated on the AVIC

database of spontaneous conversational speech. For discrimination between 4 classes of isolated Non-Verbals accuracies of 92.3% are reported using HMM as classifier. When an additional garbage class is added the accuracies drop to 80.7%, which is mainly assumed to be due to the unspecific nature of the garbage class annotations in the AVIC corpus. With SVM and HCRF similar, but slightly (approx. 2%) lower results as with HMM are observed. An additional advantage of HMM is their easy integration within a typical Automic Speech Recognition framework. MFCC and PLP based features were investigated, and lead to similar results. Addition of the extra low-level features in the HMM framework as used for static modelling did not result in any further gain.

In future works we will provide results on integrated decoding of Non-Verbals. Further, approaches for speech/Non-Verbal discrimination based on evolution of low and mid-level descriptors over time need to be investigated. I. e., tracking of voice pitch variations, loudness envelopes and rhythm of speech. Also, parameter optimisation for HMM has to be applied for each class of Non-Verbals, individually. For example, laughter is more complex than breathing, thus it requires more model parameters. Also, other modelling techniques need to be investigated such as HMM/SVM hybrids, and Long-Short-Term-Memory (LSTM) neural networks. Methods for detecting non-verbal vocalisations combined with speech must be researched. Especially laughter often occurs while a person is uttering words. It is a great challenge to detect that the person is laughing, and then detect the spoken content. Speech while laughing is quite different from regular speech regarding its acoustic properties. Also explicit methods for detection of disfluencies such as incomplete words, corrections, stuttering or repetitions must be found as it is not possible to include all combinations of incomplete words in the dictionary and the language model.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 211486 (SEMAINE).

References

[Campbell, 2007] Campbell, N.: On the use of nonverbal speech sounds in human communication. In: COST 2102 Workshop, pp. 117–128 (2007)

[Campbell et al., 2005] Campbell, N., Kashioka, H., Ohara, R.: No laughing matter. In: Proceedings of INTERSPEECH 2005, pp. 465–468 (2005)

[Decaire, 2000] Decaire, M.W.: The detection of deception via non-verbal deception cues. Law Library 1999 - 2001 (2000)

[Goto et al., 1999] Goto, M., Itou, K., Hayamizu, S.: A real-time filled pause detection system for spontaneous speech recognition. In: Eurospeech 1999, pp. 227–230 (1999)

[Gunawardana et al., 2005] Gunawardana, A., Mahajan, M., Acero, A., Platt, J.C.: Hidden conditional random fields for phone classification. In: Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH), Lisbon, Portugal (2005)

[Hermansky, 1990] Hermansky, H.: Perceptual linear predictive (plp) analysis of speech. Journal of the Acoustical Society of America 87(4), 1738–1752 (1990)

[Kennedy and Ellis, 2004] Kennedy, L.S., Ellis, D.P.W.: Laughter detection in meetings. In: NIST ICASSP 2004 Meeting Recognition Workshop, Montreal (2004)

[Knox and Mirghafori, 2007] Knox, M., Mirghafori, M.: Automatic laughter detection using neural networks. In: Proceedings of INTERSPEECH 2007 (2007)

[Kompe, 1997] Kompe, R.: Prosody in Speech Understanding Systems. Springer, Heidelberg (1997)

[Lafferty et al., 2001] Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of International Conference on Machine Learning (ICML) (2001)

[Lickley et al., 1991] Lickley, R., Shillcock, R., Bard, E.: Processing disfluent speech: How and when are disfluencies found? In: Proceedings of European Conference on Speech Technology, vol. 3, pp. 1499–1502 (1991)

[Reiter et al., 2007] Reiter, S., Schuller, B., Rigoll, G.: Hidden conditional random fields for meeting segmentation. In: Proc. ICME 2007, Beijing, China, pp. 639–642 (2007)

[Schuller et al., 2007] Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V.: The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals. In: Proc. INTERSPEECH 2007, Antwerp, Belgium, pp. 2253–2256 (2007)

[Schuller et al., 2007] Schuller, B., Müller, R., Hörnler, B., Hoethker, A., Konosu, H., Rigoll, G.: Audiovisual recognition of spontaneous interest within conversations. In: Proc. of Intern. Conf. on Multimodal Interfaces, ACM SIGHI, Nagoya, Japan, pp. 30–37 (2007)

[Schuller et al., 2008] Schuller, B., Wimmer, M., Mösenlechner, L., Kern, C., Arsic, D., Rigoll, G.: Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space? In: Proceedings of ICASSP 2008, Las Vegas, Nevada, USA (2008)

[Schultz and Rogina, 1995] Schultz, T., Rogina, I.: Acoustic and Language Modeling of Human and Nonhuman Noises for Human-to-Human Spontaneous Speech Recognition. In: Proc. ICASSP-1995, Detroit, Michigan, vol. 1, pp. 293–296 (1995)

[Truong and van Leeuwen, 2005] Truong, K.P., van Leeuwen, D.A.: Automatic detection of laughter. In: Proceedings of Interspeech, Lisbon, Portugal, pp. 485–488 (2005)

[Ward, 1991] Ward, W.: Understanding spontaneous speech: the phoenix system. In: Proceedings of ICASSP, Toronto, pp. 365–367 (1991)

[Young, 1996] Young, S.: Large vocabulary continuous recognition: review. IEEE Signal Processing Magazine 13(5), 45–57 (1996)

[Young et al., 2006] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK book (v3.4). Cambridge University Press, Cambridge (2006)