FINE-TUNING HMMS FOR NONVERBAL VOCALIZATIONS IN SPONTANEOUS SPEECH: A MULTICORPUS PERSPECTIVE

Dmytro Prylipko[1], *Björn Schuller*[2], *Andreas Wendemuth*[1]

[1]Cognitive Systems, Department of Electrical Engineering and Information Technology,
Otto von Guericke University Magdeburg, Germany
[2]Institute for Human-Machine Communication,
Department of Electrical Engineering and Information Technology,
Technische Universität München, Germany
dmytro.prylipko@ovgu.de

ABSTRACT

Phenomena like filled pauses, laughter, breathing, hesitation, etc. play significant role in everyday human-to-human conversation and have a significant influence on speech recognition accuracy [1]. Because of their nature (e. g. long duration), they should be modeled with different number of emitting states and Gaussian mixtures. In this paper we address this question and try to determine the most suitable method for finding these parameters: we provide an examination of two methods for optimization of hidden Markov model (HMM) configurations for better classification and recognition of nonverbal vocalizations within speech. Experiments were conducted on three conversational databases: TUM AVIC, Verbmobil, and SmartKom. These experiments show that with HMMs configurations tailored to a particular database we can achieve 1–3 % improvement in speech recognition accuracy with comparison to a baseline topology. An in-depth analysis of discussed methods is provided.

Index Terms— Spontaneous speech, nonverbals, laughter recognition, multiple corpora.

1. INTRODUCTION

Currently, state-of-the-art speech recognition technology provides acceptable level of performance for read speech, while there is still a large room for improvement in spontaneous speech recognition. Filled pauses, repairs, hesitations, partial words, repetitions and disfluences, included in conversational speech, complicate the recognition task. This may be crucial, as in natural and conversational speech their percentage may be surprisingly high [2]. At the same time, they enrich the spoken words content with paralinguistic information, which is vital for determining speaker's state and intention underlying the utterance [3]. An explicit modeling of nonverbal and nonspeech vocalizations can improve the recognition accuracy and provide additional information contained in these phenomena.

The first attempts to address this question have been made in 1990's. Schultz and Rogina in [4] performed a set of experiments with inclusion of acoustic models of human and nonhuman noises into a spontaneous speech recognizer. That work, however, was focused on speech recognition accuracy, while noise events were eliminated from recognition hypotheses. Also, no configuration optimization has been done, acoustic models for those kinds of events have been chosen the same way just as for phonemes.

Such an optimization has been executed in [5], where various dynamic and static classification methods for discriminating between different classes of isolated nonverbals were investigated. That work is focused on finding the best feature set and best hidden Markov model (HMM) configurations for nonverbal models from the point of view of a classification task. However, individual model optimization and incorporation into speech recognition is not dealt with in that paper.

Individual optimization of HMM configurations and feature subset for the best classification between speech and non-speech events has been also performed in [6], but the authors regard only a limited number of non-speech noises (namely filled pauses, laughter and applause) and also do not investigate the influence of such an optimization on the speech recognition accuracy.

In this work, we thus present an extensive approach towards examining of methods of HMM configurations optimization together with their evaluation within a spontaneous speech recognition system on multiple corpora in order to provide a broader understanding on their dependency of single databases.

2. CORPORA

We conducted our experiments on three spontaneous speech databases: the TUM Audio-Visual Interest Corpus (AVIC) [7] as was recently featured in the INTERSPEECH 2010 Paralinguistic Challenge [8], the SmartKom Home [9] corpus, and the Verbmobil I [10] corpus.

TUM AVIC is an English database containing human conversational speech between a product presenter and 21 diverse subjects. The total recording time for males resembles 5:14:30 h with 1 907 turns, for females total recording time resembles 5:08:00 h with 1 994 turns, respectively. The total duration of clean speech data (without presenter's phrases) is 2:17:37 h.

Verbmobil represents the data collected within the Verbmobil project[1]. For this work we used the German part of the corpus. The Verbmobil domain is negotiation, and the task to be solved by speakers is to arrange meetings and plan a trip. It consists of 1 658 spontaneous dialogs with 13 890 turns produced by 655 speakers. The total duration of the recorded speech is 33:51:42 h.

The authors acknowledge the support provided by the federal state Sachsen-Anhalt with the Graduiertenförderung (LGFG scholarship).

[1]http://verbmobil.dfki.de/

Table 1. Distribution and type of the nonverbals in the selected corpora for analysis.

# Instances	TUM AVIC	Verbmobil	SmartKom
Filled pause (*äh*)	–	3 944	74
Filled pause (*ähm*)	–	3 140	182
Filled pause (*hm*)	–	579	155
Breath	517	19 786	495
Human noise	973	1 510	852
Hesitation	1 258	467	94
Laugh	306	221	43
Throat clean	–	92	–
Swallow	–	214	–
Lip smack	–	5 156	–
Consent	360	–	–
Total	3 414	35 109	1 895

SmartKom is another German speech database collected during the SmartKom project[2]. The SmartKom Home scenario represents human conversation with an intelligent communication assistant at home. It contains data from 65 speakers and 130 recordings. The total time of recordings used in this work is 3:05:12 h.

For the experiments several categories of nonverbals were extracted with forced alignment (cf. details shown in Table 1). From all extracted nonverbals only those longer than 0.1 s were kept, since vocalizations shorter than 0.1 s are in most cases aligned in wrong way and thus are very error prone.

3. HMM CONFIGURATION OPTIMIZATION FOR IMPROVED CLASSIFICATION

In this section we consider the actual classification task. As indicated, classification is performed with a dynamic classifier, namely hidden Markov models. Except good discrimination ability, HMMs are the easiest choice for further inclusion into a speech recognizer. The feature set of ours consists of the typical 13 Mel-Frequency Cepstral Coefficients (MFCC) including the 0th coefficient and delta and acceleration regression coefficients. Features are extracted from frames of 25 ms length sampled at a rate of 10 ms. This type of features is convenient for common automatic speech recognition engines and are efficient for the nonverbals classification task [5].

Our interest in the ongoing is to determine the topologies for each HMM, which together provide the best classification accuracy with other conditions kept constant. The chosen topology parameters to vary and 'fine-tune' are the number of emitting states (*N*) and the number of Gaussian mixture components (*M*) for each state. We do not regard various state transition configuration (e. g., Bakis topology) since this was proven to be ineffective [5]. Note that we further assume that each state of a single HMM has the same number of mixture components.

While the best configurations can be found in several ways, only exhaustive search through all possible combination of all possible configurations of different models ensures finding the best solution. In this work we test models' topologies in a range of 3,$\cdots$,15 states and 1,$\cdots$,99 mixtures. Thus, the total number of different combinations is $(13 \times 99)^n$, where *n* is a number of different nonverbal models. Even with smaller range of possible configurations it makes a genuine exhaustive search computationlly infeasible, i. e., NP-hard.

[2]http://smartkom.dfki.de/

Table 2. Discrimination accuracy of two strategies for optimization w/o exhaustive search as compared to the baseline.

Configuration	TUM AVIC	Verbmobil	SmartKom
Baseline	71.94	86.14	70.4
Grid search	**74.58**	**90.28**	**72.19**
LBO	74.08	88.78	68.28

This problem can be eased by discrete optimization techniques like genetic algorithms (GA), but application of it in non-reported previous experiments to this task was not successful.

As an alternative, the search space can be reduced with fixation of the same configuration for all the models. This method was used in [5]. In such a case the number of combination equals $13 \times 99 = 1\,287$. However, it will not expose differences in the nature of single vocalizations. In the further we will reference this method as *grid search*.

Another approach consists in individual optimization of each single model with no respect to the individual classification task. In such a case only one single model is trained on the corresponding part of training data. The criterion function is an average likelihood of the observed test data obtained during the forced alignment – $P[O|\lambda]$. Due to very small values and dynamic range of the likelihood, the logarithm is usually preferred instead. Thus, the score for the model λ_n is defined as:

$$L(\lambda_n) = \Big(\sum_{t=1}^{T} log\big(P[O_t|\lambda_n]\big) \Big) \Big/ T, \qquad (1)$$

where $O_1, \cdots, O_T$ are test utterances for the corresponding nonverbal. Those models which provide the highest scores for the corresponding type of a nonverbal are then chosen to construct the overall configuration for the classification. This method is denoted in the ongoing as *likelihood-based optimization (LBO)*.

Both methods have been evaluated on the three conversational databases described in section 2. Each single evaluation was performed in a speaker-independent 3-fold cross-validation. For each database, the set of speakers was randomly split into three disjunctive parts. Note that each part includes a slightly different number of speakers in order to keep the balance of data between folds.

Classification performance of the configurations found with described approaches were compared to a baseline configuration (8 states and 8 mixtures for each model) found to be the best for MFCC features in previous work [5]. Obtained results are listed in Table 2.

As one can see from these results, best classification ability is provided with configurations found by grid search. This result is expected, since this method aims to obtain the best classification performance, while likelihood-based optimization gives us models with the highest likelihood of the test data, which does not mean the best classification performance between small number of classes due to inter-class similarities. The detailed description of configurations is provided in Tables 3 and 4.

4. EXPERIMENTS INCORPORATING LARGE VOCABULARY SPEECH RECOGNITION

However, the question about the best approach for speech recognition is still left open by the so far presented results. Obviously, the difference between classification of nonverbals and large vocabulary continuous speech recognition (LVCSR) lies in the significantly larger

Table 3. Best configurations obtained with grid search. Models share the same number of emitting states (N) and Gaussian mixtures (M).

	TUM AVIC		Verbmobil		SmartKom	
#	N	M	N	M	N	M
Baseline	8	8	8	8	8	8
Grid search	13	18	9	82	8	6

Table 4. Best configurations obtained with likelihood-based optimization, where emitting states (N) and Gaussian mixtures (M).

	TUM AVIC		Verbmobil		SmartKom	
#	N	M	N	M	N	M
Filled pause (*äh*)	–	–	9	45	8	2
Filled pause (*ähm*)	–	–	9	42	9	3
Filled pause (*hm*)	–	–	9	10	6	7
Breath	6	24	9	99	4	27
Human noise	5	41	3	72	3	41
Hesitation	13	22	5	14	4	4
Laugh	8	10	6	11	4	3
Throat clean	–	–	6	5	–	–
Swallow	–	–	4	19	–	–
Lip smack	–	–	7	99	–	–
Consent	12	5	–	–	–	–

number of classes in the case of LVCSR. In such a case, likelihood-based optimization is thus more likely to perform better. Let us thus now investigate this in more detail.

4.1. Experimental setup

The given configurations have been tested on the corresponding LVCSR tasks per corpus. Each recognizer was built with the same scheme. Acoustic modeling of regular phonemes was done with three-state context-independent left-to-right Gaussian mixture HMM models. Each HMM had three states (except short pause) and 32 mixtures. For the German databases (Verbmobil and SmartKom) we used an extended list of 47 phonemes. Also, some non-speech noises were included (technical and non-human noises, knocks, squeals and rustle). These sounds were modeled as regular phonemes (three states, 32 mixtures) since this work is focused on non-verbal events which belong to speech.

Following our previous protocol, evaluation of speech recognition performance has been performed in speaker-independent 3-fold cross-validation manner. However, on TUM AVIC we applied leave-one-speaker-out (LOSO) evaluation in order to increase the amount of training data taking into account the increased complexity given by this task.

We tested just the initial configurations of models – this means that no parameters from the previous stage were utilized for the speech recognizers. All parameters were estimated with the commonly employed Baum-Welch training procedure.

Language modeling was performed with bigrams. All kinds of non-speech and non-verbal events were modeled like usual words by applying their language model probabilities. An alternative approach would be to treat these vocalizations as silences, but this approach was previously shown to perform worse [4]. More sophisticated language modeling goes beyond the scope of this article.

Table 5. Word accuracy (and correctness) obtained with different configurations.

%	TUM AVIC	Verbmobil	SmartKom
including non-verbal vocalizations			
Baseline	23.50 (38.31)	72.21 (77.23)	51.42 (61.99)
Grid search	**24.28** (39.80)	**72.70** (78.13)	**52.54** (62.58)
LBO	24.01 (40.28)	72.56 (78.21)	50.79 (62.73)
excluding non-verbals vocalizations			
Baseline	15.46 (41.12)	76.88 (80.22)	57.50 (65.82)
Grid search	17.54 (40.94)	77.42 (80.47)	**58.51** (66.31)
LBO	**18.17** (40.91)	**77.53** (80.41)	58.19 (65.70)

4.2. Results

The results provided in Table 5 including non-verbal vocalizations show that, both methods give similar word accuracy (note that we prefer to stay with reporting on accuracies as opposed to the habit of reporting word error rates common in the field of LVCSR. This choice was made in order to be consistent with the previously reported accuracies which are the usual way to report on nonverbals.). As one can see, grid search still gives the best result from the point of view of word accuracies. It is interesting that LBO however provides better correctness, which means that it causes a larger number of word insertions. In this table, we also show results excluding non-verbal vocalizations for the final recognition result , i. e., these are modeled, but we calculate the results only for verbal events. This is interesting, as it shows benefit for individual likelihood optimization (except for the SmartKom corpus) when looking at the linguistic information.

5. DISCUSSION OF FINDINGS

The experimental results we observed on the three corpora speak for application of grid search for both: isolated nonverbal classification and conversational speech recognition. On the other hand, likelihood-based optimizationl gives us useful information about tendencies in acoustic modeling of individual nonverbals and could lead to better results exclusively for the *linguistic* content on two corpora. Although we did not discover any strong correlation in configurations between the various corpora, we can draw several conclusions as follow: First, one can see strong correlation between the optimal number of mixtures and the amount of the training data. This is especially observable on the SmartKom database, where breath and human noise occur much more often and have a proportional number of mixtures. Breath and lip smack on the Verbmobil corpus have more than 5 000 occurrences and their optimal mixture numbers are rather extreme – 99. Thus, there is no observed universal solution concerning optimal number of mixtures – this has to be determined taking into account the particular training material at hand.

However, human noise models expose some kind of similarity (a small number of states, more mixtures). Also, the optimal number of states for filled pauses is about 8–9. Yet, we cannot generalize on the optimal number of mixtures due to the reasons mentioned above.

In general, we can state that optimization of nonverbal models' topology can improve overall accuracy up to 3 % relatively. This can be obtained at reduced computational effort: Methods applied in this paper are exhaustive searches within given ranges by nature and can thus be made more efficient.

From Figure 1 we can see that the search spaces for both methods have noticeable trends towards a global maximum. Yet, the search

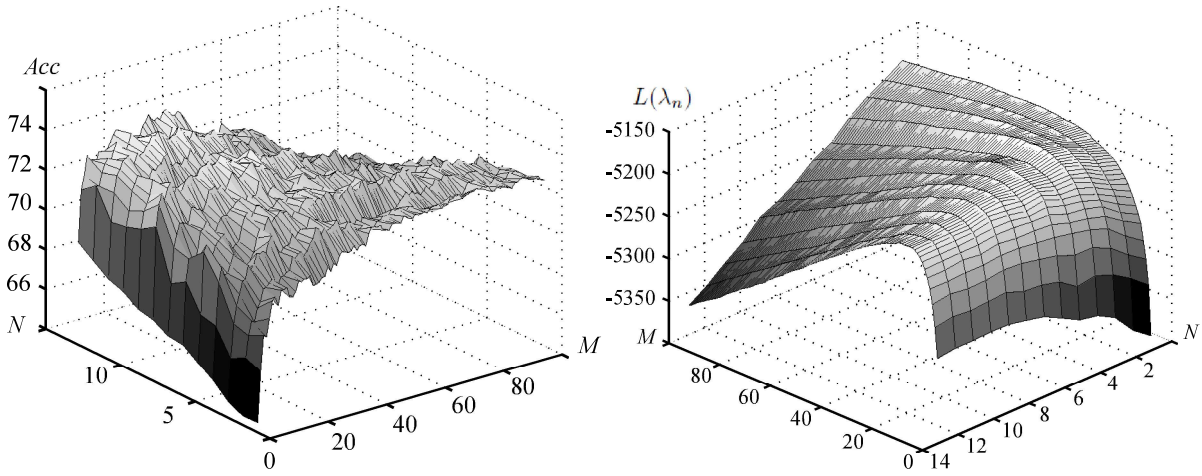

Fig. 1. Left: Search space of grid search for the TUM AVIC database: Accuracy (Acc) over emitting states (N) and Gaussian mixtures (M). Right: Search space of likelihood-based optimization for the example of *breath* on the TUM AVIC database: Average log likelihood over emitting states (N) and Gaussian mixtures (M).

space of the individual optimization is smoother at a smaller number of local maxima. This is typical for any model. This behaviour makes an application of simple gradient-based optimization algorithms reasonable, which reduces the search time dramatically.

6. CONCLUSION

We have tested two approaches for HMMs topology optimization with respect to nonverbal models. The first method performs an exhaustive search through a subset of possible models' configurations on the classification task (grid search). Within this approach all models share the same configuration. Another approach maximizes the observation likelihood of the test data for each model independently. Both methods have been first evaluated on an isolated nonverbal classification task and then on spontaneous speech recognition.

Our experimental results show that the first method provides the best configuration for both tasks. The second approach is still useful since it gives comparable results and can be easily optimized with use of more sophisticated optimization algorithms such as hill-climbing.

Fine-tuning for this class of models can increase word accuracy of speech recognizers in the range of 1–3 % respectively.

In future work we intend to investigate the method of individual HMM optimization, which maximizes not only a likelihood of observable data, but also improve the discrimination ability of the models with maximization of the margin between classes.

7. ACKNOWLEDGEMENTS

The authors acknowledge the support provided by the federal state Sachsen-Anhalt with the Graduiertenförderung (LGFG scholarship). We also acknowledge the German Research Foundation (DFG) for financing our computing cluster used for parts of this work.

8. REFERENCES

[1] E. Shriberg, "Spontaneous Speech: How Peoply Really Talk and Why Engineers Should Care," in *Proc. of EUROSPEECH*, Lisbon, Portugal, 2005, pp. 1781–1784.

[2] K. Laskowski and T. Schultz, "Detection of Laughter-in-Interaction in Multichannel Close-Talk Microphone Recordings of Meetings," in *Machine Learning for Multimodal Interaction*, LNCS 5237, pp. 149–160. Berlin-Heidelberg, 2008.

[3] N. Campbell, "On the Use of NonVerbal Speech Sounds in Human Communication," in *COST 2102 Workshop (Vietri)*. 2007, vol. 4775 LNAI, pp. 117–128, Springer Berlin-Heidelberg.

[4] T. Schultz and I. Rogina, "Acoustic and language modeling of human and nonhuman noises for human-to-human spontaneous speech recognition," in *Proc. of ICASSP'95*, Detroit, 1995, vol. 1, pp. 293–296, IEEE.

[5] B. Schuller, F. Eyben, and G. Rigoll, "Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech," in *Proceedings of the PIT'08*, Kloster Irsee, Germany, 2008, vol. LNCS 5078, pp. 99–110, Springer-Verlag.

[6] Y.-X. Li, S. Kwong, Q.-H. He, J. He, and J.-C. Yang, "Genetic algorithm based simultaneous optimization of feature subsets and hidden Markov model parameters for discrimination between speech and non-speech events," *International Journal of Speech Technology*, vol. 13, no. 2, pp. 61–73, Apr. 2010.

[7] B. Schuller, R. Müller, B. Hörnler, A. Höthker, H. Konosu, and G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations," in *Proc. of ICMI'07*, Nagoya, Japan, 2007, pp. 30–37, ACM Press.

[8] 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, pp. 2794–2797.

[9] F. Schiel, S. Steininger, and U. Türk, "The SmartKom Multimodal Corpus at BAS," in *Proc. of LREC'02*, 2002, pp. 200–206.

[10] S. Burger, K. Weilhammer, F. Schiel, and H. G. Tillmann, "Verbmobil Data Collection and Annotation," in *Verbmobil: Foundations of Speech-to-Speech Translation*, Wolfgang Wahlster, Ed., pp. 537–549. Springer, 2000.