Detection of social speech signals using adaptation of segmental HMMs

Sathish Pammi and Mohamed Chetouani

Institute for Intelligent Systems and Robotics (ISIR)
Université Pierre et Marie Curie, Paris, France
sathish.pammi@isir.upmc.fr, mohamed.chetouani@upmc.fr

Abstract

This paper proposes an approach to detect social speech signals by computing segmental features using adaptation of segmental Hidden Markov Models (HMMs). This approach uses segmental HMMs and model adaptation techniques such as Maximum Likelihood Linear Regression (MLLR) and Maximum A Posterior (MAP) in order to acquire *specific* (or adapted) segmental HMMs that are fine-tuned to detect local regions of social signals such as laughter and fillers. Several segmental features are computed on automatically segmented audio with the *specific* segmental HMMs. Subsequently, the segmental features are used to detect social signals using Support Vector Machines (SVMs). The results indicate that the proposed segmental features play a significant role in detection of social speech signals.

Index Terms: Nonlinguistic vocalizations, Detection of social signals, Hidden Markov Models, Model adaptation

1. Introduction

Despite the best efforts made over past two decades in speech recognition systems, detection of emotions and nonlinguistic vocalizations are still challenging tasks [1, 2]. Social speech signals such as laughter or fillers are most frequent vocalizations in our daily conversational speech. Several disciplines such as affective computing require tools and methods for automatically detecting social signals in speech. Traditional speech recognition frameworks have not been adequately focused on detecting nonlinguistic vocalizations such as laughs, sighs, hesitation sounds under a common and generic framework. One of the main reasons is that obtaining phonetic representation or a pronunciation dictionary for such vocalizations is an incredibly difficult task.

Schuller et al. [3, 2] show that integrating likelihood features derived from Non-negative Matrix Factorization into Bidirectional Long Short-Term Memory Recurrent Neural Networks provides better results in terms of discriminating nonlinguistic vocalizations from speech. Most of previous studies (e.g. [4, 5]) on automatic laughter detection from audio are based on frame-level acoustic features as parameters to train machine learning techniques, such as Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs). However, segmental approaches that capture higher-level events have not been adequately focused yet.

Pammi et al. [6], in a recent work, proposed a segmental approach to detect nonlinguistic vocalizations using ALISP (Automatic Language Independent Speech Processing) techniques. The main advantage of ALISP models is purely data-driven; and they can segment *any* audio signal into pseudo-phonetic units and provide corresponding segment labels. In the work, ALISP segmental models are adapted using Maximum Likelihood Linear Regression (MLLR)[7] and Maximum A Posterior (MAP)[8, 9] techniques. The resulting adapted models can then be used to detect local regions of nonlinguistic vocalizations, using the standard Viterbi algorithm.

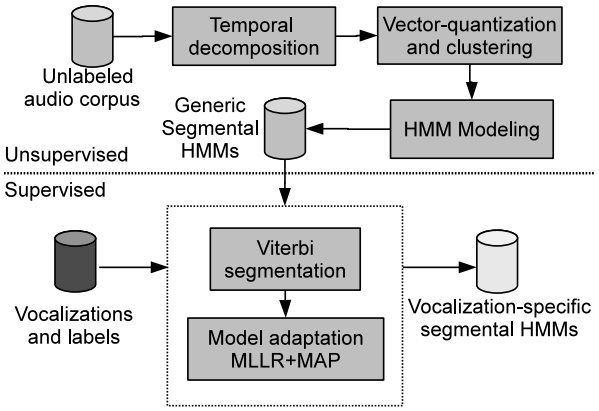

Figure 1: Workflow of two-stage methodology for acquiring *specific* segmental HMMs that are adapted for each type of social signals: (i) training of *generic* segmental HMMs using ALISP-based unsupervised techniques; (ii) acquiring *specific* segmental HMMs that are tuned for specific vocalizations using supervised acoustic model adaptation techniques.

In this paper, we extend our previous approach further to compute additional segmental-level features by adapting *generic* segmental HMMs to *vocalization specific* segmental HMMs. The paper is organized as follows: Section 2 explains our approach to extract additional segmental features using adaptation of segmental HMMs. In Section 3, we present experimental evaluation of the proposed method on social signals corpora. Conclusions follow in Section 4.

2. Approach

This section describes our generic approach to compute segmental level features using adaptation of segmental HMMs. This methodology aims at obtaining *specific* segmental HMMs that are specialized in finding similar spectral regions of target vocalizations. We can obtain such models by applying model adaptation techniques on *generic* segmental HMMs that are trained using ALISP methodology. The *specific* segmental HMMs can facilitate Viterbi decoding algorithm to detect similar spectral regions from audio.

As shown in Figure 1, the workflow of this framework can be broadly divided into two stages: (i) the training of *generic*

segmental HMMs on huge unlabeled audio corpus by using ALISP technology; (ii) the acquisition of *specific* segmental HMMs by adapting *generic* HMMs with MLLR and MAP techniques. In this approach, we intend to train vocalization specific segmental HMMs in order to find corresponding target regions in audio while decoding with Viterbi algorithm.

2.1. Generic segmental HMMs

Generic segmental HMMs are acquired using ALISP methodology [10, 11, 12]. ALISP method is an established technique to train segmental HMMs in an unsupervised approach. According to this methodology, the set of ALISP models can be automatically acquired from unlabeled audio corpus through parameterization, temporal decomposition, vector quantization, and Hidden Markov Modeling. Firstly, temporal decomposition [13] is used to obtain an initial segmentation of the audio data into quasi-stationary segments after parameterization of audio. The detailed algorithm of interpolation functions used in temporal decomposition can be found in [14]. Secondly, unsupervised clustering of initial segments is performed via Vector Quantization [15]. In order to train robust models of ALISP units on the basis of the initial segments resulting from the Temporal Decomposition, the ALISP approach uses Hidden Markov Modeling techniques. HMM training is performed using the HTK toolkit [16]. It is mainly based on Baum-Welch reestimations and an iterative procedure of refinement of the models. A dynamic split of the state mixtures is used to fix the number of Gaussians of each ALISP model. After this training step, one can obtain a set of *generic* segmental HMMs.

2.2. Specific segmental HMMs

The acquired generic models in the previous step can be used for obtaining pseudo-phonetic segments and corresponding labels. In this step, we adapt such models by providing supervised adaptation data of vocalizations. Firstly, the generic models segment the annotated audio and acquire segment labels using Viterbi decoding algorithm as shown in Figure 1. The pseudo-phonetic segmental labels, adaptation corpus and its annotation are required for the second-stage. Secondly, MLLR adaptation approach is applied to estimate a set of linear transformations for the mean and variance parameters for reducing mismatch between the initial generic segmental HMMs and the adaptation set. Finally, the model is further adapted (means, variances and transition probabilities) using MAP approach considering MLLR adapted model as prior knowledge. Therefore, adaptation of ALISP models uses MLLR followed by MAP approaches. In this way, the models are expected to deviate from each other for discriminating nonlinguistic vocalizations. Figure 1 illustrates the workflow used to obtain *vocalization specific* segmental HMMs.

Acquisition of *specific* segmental HMMs conceptually resembles a hierarchy in HMM modeling as shown in Figure 2.

2.3. Segmental feature extraction

The combination of *specific* segmental HMMs can be used for pseudo-phonetic segmentation. We compute segmental-level features on such pseudo-phonetic units that are acquired using Viterbi algorithm [17] – a well established technique for decoding an HMM sequence of states. This decoding algorithm is used in order to transform an observed sequence of speech features into a string of recognized ALISP units. In this work, a combined set of adapted ALISP models are used to discriminate

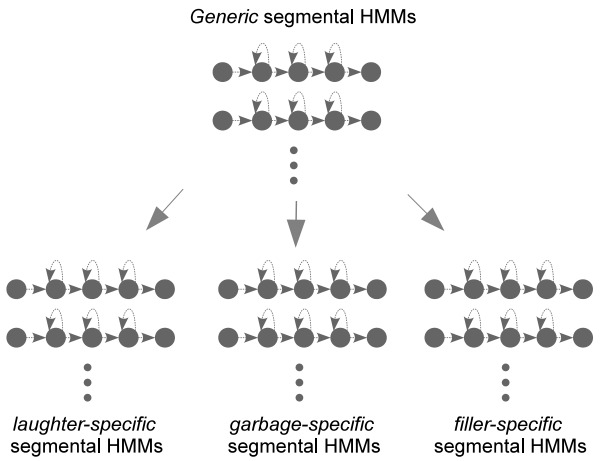

Figure 2: Hierarchy in acquisition of *specific* segmental HMMs. In this work, 33 *generic* segmental HMMs are adapted into 99 (i.e. 33 * 3 types of vocalization categories) *vocalization specific* segmental HMMs.

social vocalizations. Therefore, the labels of ALISP sequences generated from the Viterbi decoding are expected to follow a naming convention in order to support symbolic level post processing for computing segmental-level features.

The other main advantage of segmental HMMs is a possibility to operate on the level of symbols and symbolic sequences. Viterbi decoded sequence of labels contain timestamps of each segment and also corresponding maximum likelihood values. The segmental labels contain information about its hierarchy (Figure 2) of *generic* and *specific* segmental information. This information can be used as segmental cues for detection of social speech signals. In order to incorporate contextual features, we can also use a simple voting scheme that uses a sliding window on Viterbi decoded sequence to compute votes obtained for each class as a feature.

3. Experimental evaluation

In this section, we describe an experimental evaluation of the proposed method in comparison with baseline system [1] in detection of social signals.

3.1. Corpora

As explained in Section 2, this method is a two-stage methodology that requires two different corpora. In the first stage, *generic* segmental HMMs (ALISP models) are trained with approximately 240 hours of speech corpus selected from 26 days of complete broadcast audio of 13 French radio streams.

SSPNet Vocalisation Corpus (SVC) [1] is used in the second stage for supervised training in order to obtain *specific* segmental HMMs for each type of nonlinguistic vocalizations. The SVC audio corpora of social signals contains gold-standard annotations of laughter, filler and garbage. The corpus was extracted from a collection of 60 phone calls involving 120 subjects (63 female, 57 male). It contains 2763 utterances of duration about 11 seconds. Among them, we used 1583, 500, and 680 utterances for training (*train*), development (*devel*) and test (*test*) sets respectively. Overall, this corpus includes 1158 laughter (649 for *train*, 225 for *devel*, 284 for *test*), 2988 filler (1710 for *train*, 556 for *devel*, 722 for *test*) vocalizations.

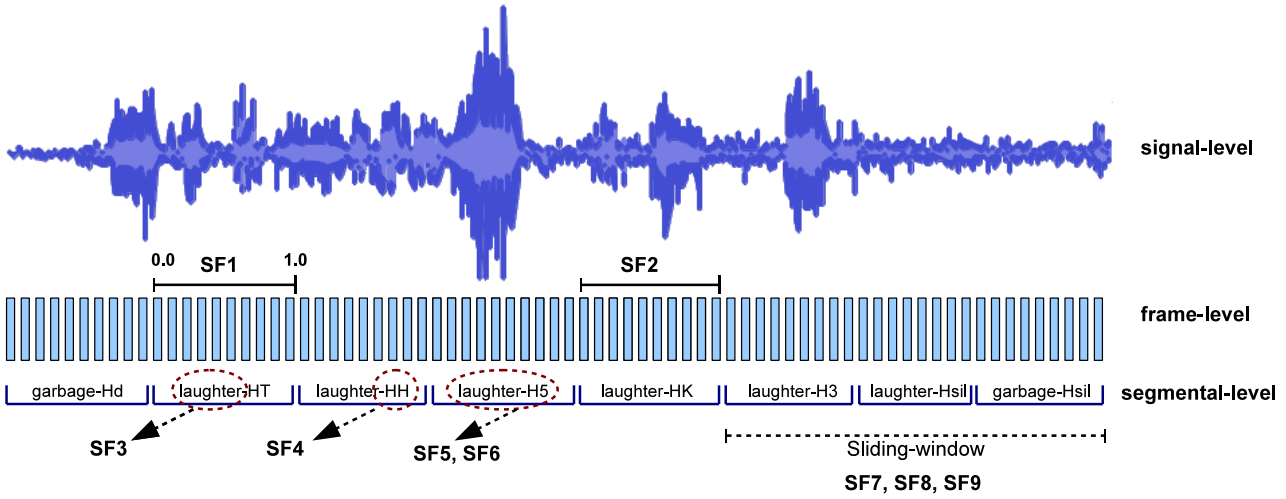

Figure 3: The segmentation is obtained by using *vocalization specific* segmental HMMs and Viterbi decoding algorithm. Segmental features (SFs) used for the detection of social speech signals are: *SF1* – normalized frame position within segment (0 <= *SF1* <= 1; i.e. 0.0 for segment's start frame; 1.0 for segment's end frame); *SF2* – number of frames in current segment; *SF3* – type-of predicted vocalization (i.e. class label); *SF4* – type-of *generic* class label (i.e. ALISP label – hierarchical information); *SF5* – segmental label predicted in Viterbi decoding; *SF6* – maximum-likelihood estimated by Viterbi algorithm; *SF7, SF8, SF9* – number of votes counted for each type of vocalization (i.e. laughter, filler, garbage);

3.2. Baseline system

As described in [1], the base line system uses 141 feature descriptors per frame. The frame-wise features include MFCCs 1–12 and logarithmic energy are computed along with their first and second order delta regression coefficients. In addition, the features also include voicing probability, HNR, F0 and zero-crossing rate, as well as their first order deltas. Then, for each frame-wise feature descriptor the arithmetic mean and standard deviation across the frame itself and eight of its neighbouring frames (four before and four after) are calculated as additional features.

The baseline system used linear kernel Support Vector Machines (SVM) / Support Vector Regression (SVR), which are known to be robust against overfitting. As training algorithm, it uses Sequential Minimal Optimisation (SMO). The results of the baseline system are shown in Table 1.

3.3. Segmental HMMs: *generic* vs. *specific*

Generic segmental HMMs were trained using ALISP methodology with 240 hours of unlabeled radio corpus. The unlabeled audio corpus has been modeled by a set of 32 ALISP segmental HMMs (i.e. pseudo-phonetic HMMs) along with a silence model. This set can be considered as an universal acoustic model because of its training database includes all possible sounds such as music, laughter, advertisements. This set of models can be used not only for segmenting any audio, but also for getting pseudo-phonetic (symbolic) transcription. For the transcription, the segmentation system uses 32 ALISP symbols (such as `HA`, `HB` and `H4`), referring each of the segmental HMMs, in addition to a silence label (`Hsil`).

In the next step, we adapted the generic ALISP segmental HMMs into vocalization specific segmental HMMs by using nonlinguistic vocalizations as adaptation data. As shown in Figure 2, *generic* segmental HMMs were adapted to vocalization specific segmental HMMs such as *laughter-specific*, *filler-specific*, and *garbage-specific* segmental HMMs. The 33 *generic* HMMs has been adapted to 99 (i.e. 33 * 3 types of vocalization categories) *vocalization specific* segmental HMMs. In order to facilitate combining the two sets, vocalization specific adapted models were renamed with its *type of* vocalization. For example, laughter-specific adapted models are renamed with `HA` to `laughter-HA`, `H4` to `laughter-H4`, and so on. The combined set of the models (i.e. set of *specific* segmental HMMs) were used to segment social signals corpus using Viterbi algorithm.

A standard set of features that are typical for automatic recognition systems have been used for HMM modeling and adaptation. The parameterization of audio data is done with Mel Frequency Cepstral Coefficients (MFCC), calculated on 20 ms windows, with a 10 ms shift. For each frame, a Hamming window has been applied and a cepstral vector of dimension 15 was computed and appended with first order deltas.

3.4. Extraction of segmental cues

We computed segmental features for social speech signals as shown in Figure 3. The features from *SF1* to *SF6* are explained in the figure description. In order to take account of symbolic level contextual information, we compute features from *SF7* to *SF9* by counting the number of votes for each class descriptor. As described in Section 2.3, a sliding window counts 'yes/no' votes depending on whether symbols in its range belong to target vocalization. The sliding window size can be 3 and/or 5. For example, as shown in Figure 3, a sliding window of size 3 is used to compute the number of votes for each category; where *laughter*, *filler*, and *garbage* get 2, 0, and 1 votes respectively. In this work, we calculated such features using two sliding windows: one of its size 3; and another of its size 5.

3.5. Results and discussion

In order to understand the influence of segmental features, we trained with the same training algorithm, SVMs, used in the baseline system. *C* indicates SVM's complexity parameter.

Table 1 shows the results of detection of social speech signals. Initially, we trained SVMs with segmental features alone that are explained in Section 3.4. The Unweighted Average of Area Under Curve (UAAUC) measures [18] in detecting non-linguistic vocalizations on development and test set are 90.9% and 86.73% respectively. When compared to the baseline system, we found at least 3% better performance using segmental features alone. Later, we also trained with a combined set of features (segmental features + baseline features). The performance is observed as 92.50% and 88.15% on development and test sets respectively in terms of UAAUC measures. Therefore, the system yields a consistent increase of 4.9% and 4.85% on the UAAUC measure when compared to baseline performance on development and test sets respectively.

[%]	C	Devel	Test
Baseline features only [1]			
AUC[Laughter]	0.1	86.2%	82.9%
AUC[Filler]	0.1	89.0%	83.6%
UAAUC		87.6%	83.3%
Segmental features only			
AUC[Laughter]	0.1	90.90%	88.44%
AUC[Filler]	0.1	90.90%	85.02%
UAAUC		90.90%	86.73%
Segmental features + baseline features			
AUC[Laughter]	0.1	92.60%	89.74%
AUC[Filler]	0.1	92.40%	86.55%
UAAUC		92.50%	**88.15%**

Table 1: Results on detection of laughter and fillers in audio

The segmental features show a clear positive impact on the performance of detection of nonlinguistic vocalizations. Interestingly, the acoustic modeling in HMMs and model adaptation uses a standard set of automatic speech recognition features (i.e. MFCCs and their deltas), and the segmental features alone performed at least 3% better than baseline system.

4. Conclusion

We described a data-driven approach in detection of social speech signals using adaptation of segmental HMMs. We used unsupervised ALISP methodology to obtain generic segmental HMMs. Then, we adapted generic segmental HMMs to vocalization specific segmental HMMs using MLLR and MAP supervised adaptation techniques.

In this paper, we mainly focused on extraction of additional segmental features that contain temporal structure of spectral distribution of social speech signals such as laughter and fillers. We computed several frame-wise segmental features and symbolic-level contextual features for detection of the social speech signals. When compared to the baseline system, the results consistently indicate that: the described segmental features alone yielded at least 3% better performance in terms of UAAUC; and a combined set of features that include baseline features yielded around 4.9% of increase in UAAUC.

5. Acknowledgements

This work was supported by the European Union Seventh Framework Programme under grant agreement n288241 through the Michelangelo project. This work is partially supported by AVATAR 1.1 project. Thanks to Houssem Khemiri for providing trained ALISP models as part of CNRS project collaboration.

6. References

[1] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech 2013*, 2013.

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

[3] B. Schuller and F. Weninger, "Discrimination of speech and non-linguistic vocalizations by non-negative matrix factorization," in *Proceedings of ICASSP 2010*, 2010, pp. 5054–5057.

[4] K. Truong and D. Van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Communication*, vol. 49, no. 2, pp. 144–158, 2007.

[5] M. Knox and N. Mirghafori, "Automatic laughter detection using neural networks," in *Proceedings of INTERSPEECH*, 2007, pp. 2973–2976.

[6] S. Pammi, H. Khemiri, D. Petrovska-Delacretaz, and G. Chollet, "Detection of nonlinguistic vocalizations using alisp sequencing," in *Proceedings of ICASSP 2013*, 2013.

[7] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer speech and language*, vol. 9, no. 2, p. 171, 1995.

[8] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 2, pp. 291–298, 1994.

[9] C. Chesta, O. Siohan, and C. Lee, "Maximum a posteriori linear regression for hidden markov model adaptation," in *Proc. EuroSpeech*, vol. 1, 1999, pp. 211–214.

[10] G. Chollet, J. Cernocký, A. Constantinescu, S. Deligne, and F. Bimbot, *Towards ALISP: a proposal for Automatic Language Independent Speech Processing*, ser. NATO ASI Series. Springer Verlag, 1999, pp. 357–358.

[11] A. El Hannani, D. Petrovska-Delacrétaz, B. Fauve, A. Mayoue, J. Mason, J. F. Bonastre, and G. Chollet, "Text independent speaker verification." in *Guide to Biometric Reference Systems and Performance Evaluation*. Springer Verlag, 2009.

[12] M. Padellini, F. Capman, and G. Baudoin, "Very low bit rate (vlbr) speech coding around 500 bits/sec," in *EUSIPCO*, 2004.

[13] B. Atal, "Efficient coding of lpc parameters by temporal decomposition," in *in Proceedings of ICASSP 1983*, April 1983, pp. 81–84.

[14] F. Bimbot, "An evaluation of temporal decomposition," Acoustic Research Department AT&T Bell Labs, Tech. Rep., 1990.

[15] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communication*, vol. 28, no. 1, pp. 84–95, January 1980.

[16] *Cambridge University Engineering Department. HTK: Hidden Markov Model ToolKit, http://htk.eng.cam.ac.uk.*

[17] S. Young, N. Russell, and J. Thornton, "Token passing: a conceptual model for connected speech recognition systems," Cambridge University, Tech. Rep., 1989.

[18] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.