DETECTION OF NONLINGUISTIC VOCALIZATIONS USING ALISP SEQUENCING

Sathish Pammi[1], *Houssemeddine Khemiri*[1,2], *Dijana Petrovska-Delacrétaz*[2], *Gérard Chollet*[1,3]

[1] Institut Mines-Télécom ; Télécom ParisTech ; CNRS LTCI
[2] Institut Mines-Télécom ; Télécom SudParis ; CNRS SAMOVAR
[3] Department of Electrical & Computer Engineering, Boise State University, Idaho, USA
pammi,khemiri,chollet@telecom-paristech.fr,dijana.petrovska@it-sudparis.eu

ABSTRACT

In this paper, we present a generic methodology to detect nonlinguistic vocalizations using ALISP (Automatic Language Independent Speech Processing), which is a data-driven audio segmentation approach. Using Maximum Likelihood Linear Regression (MLLR) and Maximum A Posterior (MAP) techniques, the proposed method adapts ALISP models, which then facilitate detection of local regions of nonlinguistic vocalizations with the standard Viterbi decoding algorithm. We also illustrate how a simple majority voting scheme, using a sliding window on ALISP sequences, can be helpful in eliminating outliers from the Viterbi-predicted sequence automatically. We evaluate the performance of our method on detection of laughter, a nonlinguistic vocalization, in comparison with global acoustic models such as GMMs, left-to-right HMMs and ergodic HMMs. The results indicate that adapted ALISP acoustic models perform better than global acoustic models in terms of F-measure. Moreover, our majority voting scheme on ALISP-sequences further improves the performance yielding, in total, an increase of 19.6%, 8.1% and 5.6% on the F-measure against global acoustic models GMMs, left-to-right HMMs, and ergodic HMMs respectively.

Index Terms— ALISP sequencing, acoustic models, audio segmentation, model adaptation

1. INTRODUCTION

Despite the best efforts made over past two decades in speech recognition systems, detection of nonlinguistic vocalizations such as laughter, sighs, breathing, or hesitation sounds is still a challenging task [1]. Such vocalizations are most frequent vocalizations in our daily conversational speech. Detection of the presence of these vocalizations is useful in several disciplines; for example, in affective computing. Moreover, automatic speech recognition systems also require detection of nonlinguistic vocalizations to improve the performance. Traditional speech recognition frameworks have not been adequately focussed 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.

Laughter is one of the complex nonlinguistic vocalizations [2] that conveys a wide range of messages with different meanings. Most of previous studies (e.g. [3, 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). Schuller et al. [6, 1] 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. However, segmental approaches that capture higher-level events have not been adequately focussed due to the nonlinguistic nature of laughter.

In this paper, we present a generic framework to detect nonlinguistic vocalizations using ALISP-based approaches [7], which have been successfully applied for speaker verification [8], or low bit-rate coding [9]. The main advantages of these approaches are not only purely data-driven, but also they can segment *any* audio signal into pseudo-phonetic units and provide corresponding segment labels, referred to as 'ALISP sequencing'. Our method adapts the ALISP segmental models using Maximum Likelihood Linear Regression (MLLR)[10] and Maximum A Posterior (MAP)[11, 12] techniques. The resulting adapted models can then be used to detect local regions of nonlinguistic vocalizations, using the standard Viterbi algorithm. Experiments on a laughter-annotated audio corpus show the usefulness of the proposed method.

The paper is organized as follows: Section 2 explains the proposed methodology to detect any type of nonlinguistic vocalizations, while in Section 3, experimental evaluation of the proposed method, on an laughter-annotated corpus, is presented. Conclusions follow in Section 4.

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 270780. The second author is financially supported by the ANR-SurfOnHertz project.

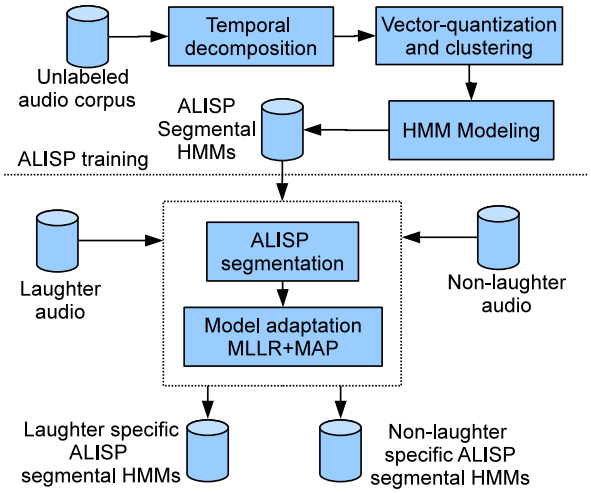

Fig. 1. Workflow of the proposed methodology for ALISP-based acoustic model adaptation to detect nonlinguistic vocalizations ('Laughter' is used as an example for a specific set of nonlinguistic vocalizations)

2. METHODOLOGY

This section describes our generic framework on detection of nonlinguistic vocalizations using ALISP sequencing. The main intension behind the proposed methodology is to adapt ALISP segmental HMMs in order to facilitate Viterbi decoding algorithm to detect similar regions from audio. The workflow of the framework can be broadly divided into two stages: (i) training ALISP models on huge unlabeled audio corpus; (ii) adaptation of ALISP models using MLLR and MAP approaches. This is illustrated in Figure 1, which shows the workflow of the proposed methodology for the specific example of detecting laughter vocalizations from audio. Laughter vocalizations are used as adaptation data to model laughter specific segmental HMMs, while nonlaughter audio (i.e. audio excluding laughter vocalizations) is used for getting non-laughter specific segmental HMMs. Finally, a combined set of HMMs are used to discriminate laughter from audio with the help of Viterbi decoding algorithm.

2.1. ALISP training

ALISP training is an established technique to train segmental HMMs in an unsupervised approach. As explained in [7, 8, 9], the set of ALISP models is automatically acquired from unlabeled audio corpus through parameterization, temporal decomposition, vector quantization, and Hidden Markov Modeling. This set of ALISP models can be used to transform a new incoming audio data in a sequence of ALISP symbols.

After the parameterization step, temporal decomposition is used to obtain an initial segmentation of the audio data into quasi-stationary segments. This method was introduced originally by Atal [13] as a nonuniform sampling and interpolation procedure for efficient parameter coding. The detailed algorithm to find interpolation functions can be found in [14].

The next step in the ALISP process is the unsupervised clustering procedure performed via Vector Quantization [15]. This method maps the P-dimensional vector of each segment provided by the temporal decomposition into a finite set of L vectors. Each vector is called a code vector or a codeword and the set of all the codewords is called a codebook. The codebook size L defines the number of ALISP units.

The final step is performed with the Hidden Markov Modeling procedure. The objective here is to train robust models of ALISP units on the basis of the initial segments resulting from the Temporal Decomposition and Vector Quantization steps. HMM training is performed using the HTK toolkit [16]. It is mainly based on Baum-Welch reestimations and on 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 is over, one obtains a set of ALISP segmental HMMs.

2.2. ALISP segmentation and model adaptation

The acquired ALISP models, in the previous step, can be used for pseudo-phonetic sequencing. In the current step, we adapt ALISP models for detecting local regions of nonlinguistic vocalizations by providing some supervised adaptation data. Firstly, ALISP models segment the adaptation data and acquire segment labels as shown in Figure 1. Next, using the segment labels and adaptation data, MLLR adaptation approach is applied to estimate a set of linear transformations for the mean and variance parameters for reducing mismatch between the initial ALISP models and the adaptation set. Finally, the model is further adapted using MAP approach considering MLLR adapted model as a prior knowledge. Therefore, adaptation of ALISP models uses MLLR followed by MAP approaches.

We propose to adapt ALISP models for specific nonlinguistic vocalizations that need to be detected as well as for the remaining data excluding the vocalizations. In this way, the models are expected to deviate from each other in discriminating nonlinguistic vocalizations from speech. Figure 1 considers laughter as one of the nonlinguistic vocalizations. As shown in the figure, the adaptation is performed on the annotated laughter vocalizations as well as on the nonlaughter part of audio corpora excluding laughter vocalizations.

2.3. Viterbi decoding and symbolic-level smoothing

The Viterbi algorithm [17], a well-established technique for decoding an HMM sequence of states, is used in order to transform an observed sequence of speech features into a string of recognized ALSIP units. In this work, a combined

set of adapted ALISP models are used to discriminate nonlinguistic vocalizations from speech. Therefore, the labels of ALISP sequences that are generated from the Viterbi decoding are expected to follow a naming convention in order to support symbolic level post processing.

The other main advantage of segmental HMMs is a possibility to operate on the level of symbols and sequences. The outliers in the Viterbi decoded sequence can be post-processed using contextual label information. This method proposes a simple voting scheme that uses a sliding window on an ALISP sequence to eliminate outliers in Viterbi-predicted sequence automatically. The sliding window counts 'yes/no' votes depending on whether or not a symbol belongs to target vocalization. The window length is always expected to be an odd number and the result of majority votes decides if the middle segment is a part of nonlinguistic vocalization.

3. EXPERIMENTAL EVALUATION

In this section, we describe an experimental evaluation of the proposed method when compared to global acoustic models in discriminating laughter from speech. Firstly, we describe the laughter-annotated experimental corpus and features used for the experimentation. Secondly, we model global HMMs (i.e. laughter versus nonlaughter models) as well as segmental HMMs by adaption of ALISP models, as described in Section 2, on laughter and nonlaughter training datasets. In addition, a combined set of laughter and nonlaughter ALISP segmental HMMs are used together to segment the test data set using the Viterbi algorithm. Consequently, the symbolic-level smoothing is applied to eliminate outliers from the predicted ALISP sequences. Finally, the results of our method are analyzed.

3.1. Experimental corpus and features

As explained in Section 2, this method is a two-stage methodology that requires two different corpuses. In the first stage, ALISP model training is done with approximately 240 hours of speech corpus selected from 26 days of complete broadcast audio of 13 French radio streams. The second stage requires supervised training material for nonlinguistic vocalizations that has manual annotation. We used a combined audio corpus that contains gold-standard laughter annotations from three different sources SEMAINE-DB [18], AVLaughterCycle [19], and Mahnob laughter databases [20]. The corpus is an appropriate mix of hilarious and conversational laughter vocalizations. The data is uniformly divided into approximately 80% for training and 20% for testing. Table 1 shows the size of laughter and nonlaughter audio (in seconds) used for training and testing.

	Laughter [sec]	nonlaughter [sec]
Training	3943	4957
Test set	853	1206
Total	4796	6163

Table 1. Training and test data sets used for experimentation

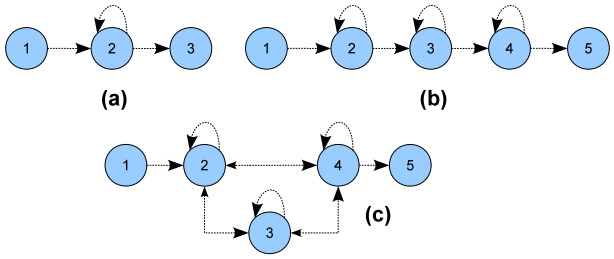

Fig. 2. Global HMM topologies: (a) Simple GMM; (b) Serial (left-to-right) HMM; (c) Ergodic (fully-connected) HMM.

A standard set of features that are typical for automatic recognition systems have been used throughout this work in order to facilitate a fair comparison among different approaches. 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 is applied and a cepstral vector of dimension 15 is computed and appended with first order deltas.

3.2. Global acoustic models vs. Adapted ALISP models

In order to detect laughter vocalizations from speech, we have trained global acoustic models such as GMMs, serial HMMs and ergodic HMMs with different HMM topologies, as shown in Figure 2. All of the above global acoustic models include an additional silence model.

ALISP segmentation models were trained with 240 hours of unlabeled radio corpus. In this work, the unlabeled audio corpus is 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 like music, laughter, advertisements etc. This set of models can be used not only for segmenting any audio, but also for getting pseudo-phonetic (symbolic) transcription. In order to represent ALISP segments, 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`). Figure 3 shows an example of the segmentation task performed by the ALISP segmental HMMs on an unseen laughter vocalization.

In the next step, we adapt the generic ALISP segmental HMMs into: (i) laughter specific ALISP segmental HMMs by using laughter vocalizations as adaptation data; (ii) nonlaughter specific ALISP segmental HMMs considering nonlaughter vocalizations (audio excluding laughter vocalizations) as adaptation data. In order to facilitate combining the two sets, laughter-specific adapted models are renamed such that `HA` to

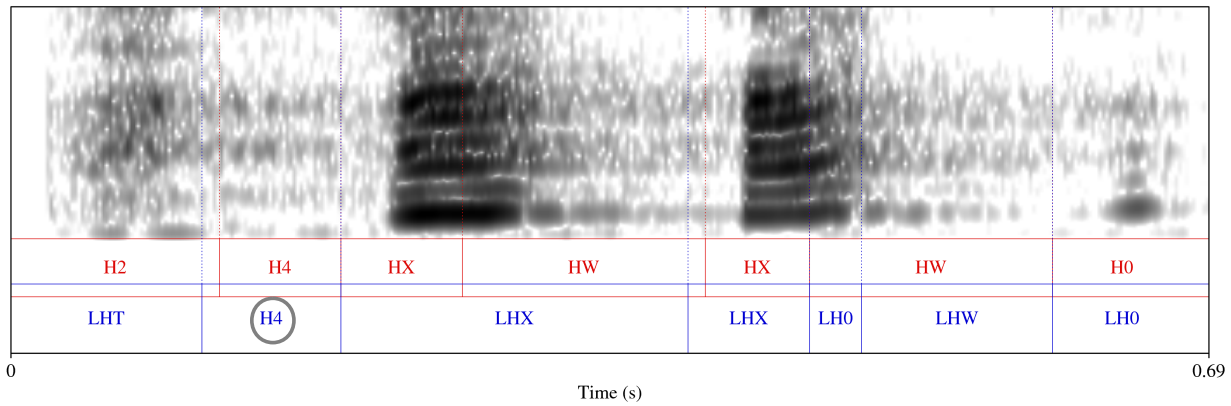

Fig. 3. Segmentation task performed on an unseen laughter vocalization by: (i) generic ALISP HMMs before model adaptation (top row labels that are in Red); (ii) Combined set of specific (or adapted) ALISP HMMs after MLLR+MAP adaptation (i.e. *ALISP-adapt*) (bottom row labels that are in Blue). The marked symbol with a circle is an outlier which can be automatically found using proposed smoothing scheme on ALISP sequences.

`LHA`, `H4` to `LH4`, and so on. On the other hand, nonlaughter specific adapted models keeps the same names such as `HA`, `H4`, `HB`, etc. The combined set of the models (say *ALISP-adapt*) were used to discriminate local regions of laughter. As shown in Figure 3, laughter specific regions seemed to be detected by the model except some outliers. In order to eliminate these outliers a majority voting scheme has been proposed in Section 2.3. We experimented the smoothing scheme using sliding window size 3 (*ALISP-adapt-sm3*) and 5 (*ALISP-adapt-sm5*). According to the scheme, for example, the outlier (`H4`) in Figure 3 get majority 'yes' votes in case of laughter detection if sliding window size is either 3 or 5. Such a way, we can automatically detect and eliminate the outliers.

3.3. Results and discussion

Table 2 shows the precision, recall and F-measures obtained from different approaches to detect laughter on test set. Among the global acoustic models, ergodic HMMs performed better than GMMs and serial (left-to-right) HMMs; ergodic HMMs showed high precision (92.8%) in locating laughter regions, whereas serial HMMs were relatively good in recall (86.3%) rates. When compared with adapted ALISP segmental HMMs (*ALISP-adapt*), global ergodic HMMs are still 4.2% better in precision. However, the segmental HMMs (*ALISP-adapt*) still performed better in terms of overall accuracy (F-measure) when compared to global HMMs.

Adapted ALISP HMMs provided an additional flexibility to find outliers with the help of a simple majority voting scheme. Therefore, *ALISP-adapt-sm3* and *ALISP-adapt-sm5* showed improvement in terms of F-measure when compared to *ALISP-adapt* by 2.9% and 4.4% of respectively. Overall, *ALISP-adapt-sm5* showed 94.3% precision and 93.9% recall rates and performed relatively better than all other approaches experimented in this work.

[%]	Precision	Recall	F-measure
GMMs	70.8	78.6	74.5
Serial HMMs	85.7	86.3	86.0
Ergodic HMMs	92.8	84.5	88.5
ALISP-adapt	88.6	90.9	89.7
ALISP-adapt-sm3	92.4	92.7	92.6
ALISP-adapt-sm5	**94.3**	**93.9**	**94.1**

Table 2. Frame-wise laughter detection results on the test set

4. CONCLUSION

In this paper, we proposed a generic approach for detecting nonlinguistic vocalizations using ALISP sequencing. In fact, this is the *first* time that segmental approaches are deployed for detection of nonlinguistic vocalizations. We evaluated the proposed methodology against global acoustic models such as GMMs, left-to-right HMMs and ergodic HMMs on a laughter-annotated audio corpus. We also used a standard set of features (i.e. MFCCs and deltas of MFCCs) that are typical in traditional systems. The results show that the proposed methodology yields an increase of 19.6%, 8.1% and 5.6% on F-measure against the three methods compared respectively.

With this work, we argue that the adaptation of the set of ALISP HMMs is useful in detecting local regions of nonlinguistic vocalizations. The segmental approach has further facilitated us to improve the performance using symbolic-level smoothing such as majority voting scheme with a sliding window approach.

5. REFERENCES

[1] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5840–5843.

[2] N. Campbell, H. Kashioka, and R. Ohara, "No laughing matter," in *Proceedings of the Ninth European Conference on Speech Communication and Technology (Interspeech)*, 2005, pp. 465–468.

[3] K.P. Truong and D.A. Van Leeuwen, "Automatic detection of laughter," in *Proc. Interspeech Euro. Conf*, 2005, pp. 485–488.

[4] K.P. Truong and D.A. 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] B. Schuller and F. Weninger, "Discrimination of speech and non-linguistic vocalizations by non-negative matrix factorization," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 5054–5057.

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

[8] 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.

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

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

[11] J.L. Gauvain and C.H. 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.

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

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

[14] F. Bimbot, "An evaluation of temporal decomposition," Tech. Rep., Acoustic Research Department AT&T Bell Labs, 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.H. Russell, and J.H.S Thornton, "Token passing: a conceptual model for connected speech recognition systems," Tech. Rep., Cambridge University, 1989.

[18] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: annotated multimodal records of emotionally colored conversations between a person and a limited agent," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 5–17, 2012.

[19] J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner, "The avlaughtercycle database," in *Proc. of LREC*, 2010.

[20] S. Petridis, B. Martinez, and M. Pantic, "The mahnob laughter database," *Image and Vision Computing*, 2012.