Laughter detection using ALISP-based N-Gram models

Sathish Pammi, Houssemeddine Khemiri and Gérard Chollet
Telecom ParisTech, Rue Dareau, 37-39, 75014 Paris, France
{firstname.lastname}@telecom-paristech.fr

Laughter is a very complex behavior that communicates a wide range of messages with different meanings. It is highly dependent on social and interpersonal attributes. Most of the previous works (e.g. [1, 2]) on automatic laughter detection from audio uses frame-level acoustic features as parameters to train their machine learning techniques, such as Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs) etc. However, segmental approaches that capture higher-level events have not been adequately focussed due to the nonlinguistic nature of laughter. This paper is an attempt to detect laughter regions with the help of automatically acquired acoustic segments using Automatic Language Independent Speech Processing (ALISP) [3, 4] models.

Method

The ALISP tools provide a general framework for creating speech units with little or no supervision. As shown in the Figure 1, the ALISP models are estimated on an audio database through parametrization, temporal decomposition, vector quantization, and Hidden Markov Modeling (HMM). ALISP units/segments are automatically acquired (i.e. unsupervised) segmental units from the ALISP models.

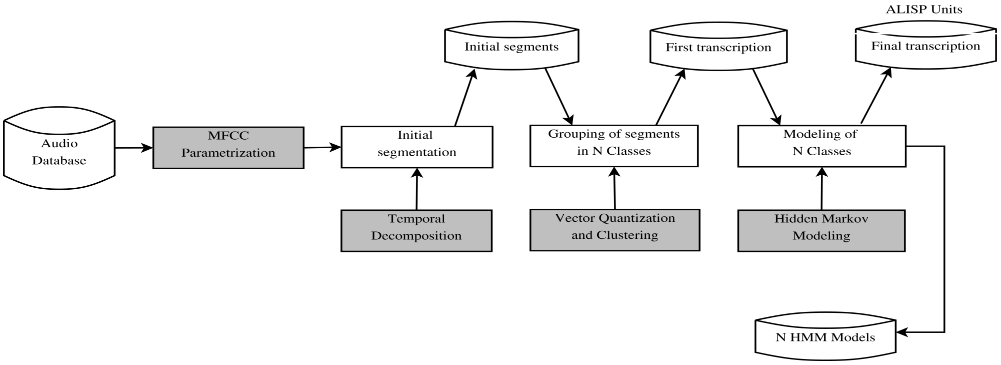

Figure 1: ALISP units acquisition and their HMM modeling

This work uses an ALISP-based automatic segmentation system which is modeled with 26 days of complete broadcast audio of 13 French radio stations provided by YACAST. This model can be considered as an universal acoustic model because of its training database includes all possible sounds like music, laughter, advertisements etc. The advantage of these models is not only the capability of segmenting any audio, but also providing appropriate symbolic level annotation for the segments. In order to represent ALISP units, the segmentation system uses 64 ALISP symbols (such as *'Ha', 'Hv'* and *'H@'*) in addition to a silence label. Figure 2 is an example of laughter audio segmented by ALISP models.

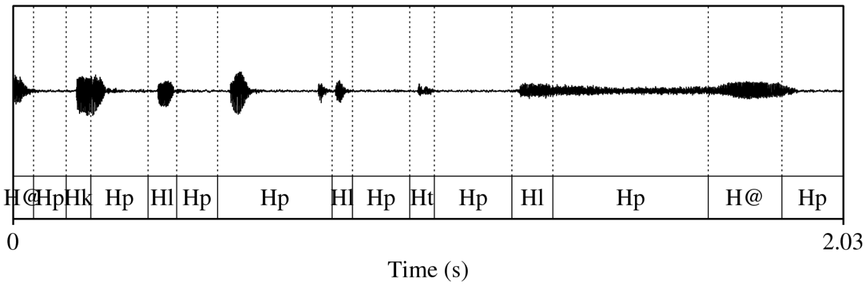

Figure 2: Laughter audio segmented by ALISP models

We hypothesize that the sequence of ALISP symbols contains the patterns of laughter. N-gram models (e.g. the sequence 'Hp-H@-Hp' is a 3-gram) on the ALISP symbolic sequence could model the patterns to detect laughter regions. A tool has been built to detect laughter from audio using linearly combined estimate of N-gram models of increasing order as follows:

$$\hat{P}(w_n \mid w_{n-2}, w_{n-1}) = \lambda_1 P(w_n \mid w_{n-2}, w_{n-1}) + \lambda_2 P(w_n \mid w_{n-1}) + \lambda_3 P(w_n)$$

Where: $\sum_i \lambda_i = 1$

In the above equation, trigram (i.e. N=3) models are mixed with bigram and unigram models. The linear interpolation of N-gram models ensure that the models suffer less from sparseness.

Evaluation and results

The ALISP-based N-gram models are trained on SEMAINE-DB [5] and AVLaughterCycle [6] databases, and the models are evaluated with Mahnob laughter database [7]. All of the three databases have manual annotations of laughter cycles. The SEMAINE-DB contains 5015 and 389 seconds of non-laughter and conversational laughter audio respectively; whereas the AVLaughterCycle DB has 3477 seconds of hilarious laughter. The MAHNOB laughter database contains 1837 and 2307 seconds of laughter and non-laughter audio respectively. As shown in Figure 3, we compared ALISP-based N-Gram models with acoustic models like GMMs, sequential (left-to-right) HMMs and ergodic (fully-connected) HMMs trained to discriminate laughter and non-laughter audio. Simple GMMs performed better precision when compared to HMMs, while ergodic HMMs provides high recall rate (93%) than GMMs. ALISP-based N-Gram models have good precision in detecting laughter, though, the recall rate is low. For example, the interpolated 5-Gram ALISP model showed more than 90% precision which indicates minimum manual intervention to find false alarms while extracting laughter from naturalistic audio resources such as radio broadcasting.

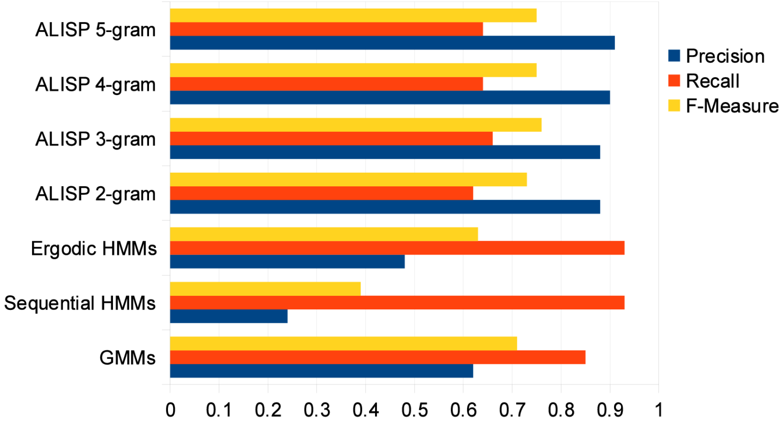

Figure 3: Performance of ALISP-based N-gram models versus GMM and HMM-based acoustic models

Discussion

The performance of ALISP-based N-grams models can be improved with more laughter training material. The ALISP symbols could be assumed as descriptions of 'very short acoustic acts'. The sequence of such cues could preserve the behavioral patterns of not only laughter, but also any other interactional vocalizations that are nonlinguistic in nature. We plan to investigate possibilities combine frame-level acoustic features with segmental features to improve the performance of laughter detection.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 270780 (ILHAIRE project). We would like to thank Prof. Dijana Petrovska for useful discussions.

References

[1] Truong, K.P. and Van Leeuwen, D.A. (2007), “Automatic discrimination between laughter and speech”, journal of Speech Communication.
[2] Knox, M. and Mirghafori, N. (2007), “Automatic laughter detection using neural networks”, proceedings of INTERSPEECH 2007.
[3] Chollet G, Cernocký J, Constantinescu A, Deligne S, Bimbot F (1999), “Towards ALISP: a proposal for Automatic Language Independent Speech Processing”, NATO ASI Series. Springer, pp 375-387.
[4] Khemiri H, and Chollet G, and Petrovska-Delacrétaz D (2011), “Automatic detection of known advertisements in radio broadcast with data-driven ALISP transcriptions”, 9th International Workshop on Content-Based Multimedia Indexing (CBMI), pp 223 – 228.
[5] McKeown, G. and Valstar, M.F. and Cowie, R. and Pantic, M. (2010), “The SEMAINE corpus of emotionally coloured character interactions”, IEEE International Conference on Multimedia and Expo (ICME), pp 1079–1084.
[6] Urbain, J. and Niewiadomski, R. and Bevacqua, E. and Dutoit, T. and Moinet, A. and Pelachaud, C. and Picart, B. and Tilmanne, J. and Wagner, J., (2010) “AVLaughterCycle: Enabling a virtual agent to join in laughing with a conversational partner using a similarity-driven audiovisual laughter animation”, Journal on Multimodal User Interfaces, pp. 47 – 58.
[7] Petridis S, Martinez B, Pantic M (nd.), “MAHNOB-Laughter Database”, Retrieved from: http://ibug.doc.ic.ac.uk/research/mahnob- laughter-database.