ALISP-based Data Compression for Generic Audio Indexing-Ver24Jan2014-VFg

Houssemeddine Khemiri, Dijana Petrovska-Delacrétaz, Gérard Chollet

ALISP-based Data Compression for Generic Audio Indexing-Ver24Jan2014-VF

Houssemeddine Khemiri*, Dijana Petrovska-Delacrétaz*, and Gérard Chollet†

*Institut Mines-Télécom
Télécom SudParis
CNRS SAMOVAR
9, rue Charles Fourier
Evry, 91130, France
dijana.petrovska@telecom-sudparis.eu
houssemeddine.khemiri@telecom-sudparis.eu

†Institut Mines-Télécom
Télécom ParisTech
CNRS LTCI
37-39, rue Dareau
Paris, 75014, France
chollet@telecom-paristech.fr

Abstract

In this paper we propose a generic framework to index and retrieve audio. In this framework, audio data is transformed into a sequence of symbols using the ALISP tools. In such a way the audio data is represented in a compact way. Then an approximate matching algorithm inspired from the BLAST technique is exploited to retrieve the majority of audio items that could be present in radio stream. The evaluations of the proposed systems are done on a private radio broadcast database provided by YACAST and other publicly available corpora. The experimental results show an excellent performance in audio identification (for advertisement and songs), audio motif discovery (for advertisement and songs), speaker diarization and laughter detection. Moreover, the ALISP-based system has obtained the best results in ETAPE 2011 (Evaluations en Traitement Automatique de la Parole) evaluation campaign for the speaker diarization task.

Introduction

For many decades, audio processing technologies have simplified the storage and accessibility to data. Actually, millions of audio documents are listened to and hundreds of them are treated every day. There are some existing systems being developed to automatically analyze and summarize audio content for indexing and retrieval purposes. Within these systems audio data are treated differently depending on the applications. For example, audio identification and audio motif discovery systems are generally based on audio fingerprinting [23] [4]. While speaker diarization systems are using cepstral features and machine learning techniques [3]. The diversity of the audio indexing techniques makes unsuitable the simultaneous treatment of audio streams when different types of audio content coexist. For example in radio streams, many types of audio data are found. These data are usually related to songs, commercials, jingles, speech or nonlinguistic vocalizations (such as laughter, sights and coughs). Therefore, a generic framework for radio broadcast indexing and retrieval is proposed.

In this paper we report our recent efforts in extending the ALISP (Automatic Language Independent Speech Processing) approach developed for speech [7] as a

generic method for radio broadcast indexing and retrieval. ALISP is a data-driven technique that was first developed for very low bit-rate speech coding [6], and then successfully adapted for other tasks such as speaker verification [8] and forgery [19], and language identification [21]. The particularity of ALISP tools is that no textual transcriptions are needed during the learning step, and only raw audio data is sufficient to train the Hidden Markov ALISP models. Moreover, the ALISP sequence of symbols is a compact representation of audio data which is exploited to index and retrieve the majority of audio items that can occur in radio streams. These items are usually: music, commercials, jingles, speech or nonlinguistic vocalizations (laughter, cough, sight,...). The proposed framework is applied on the following audio indexing tasks:

- Audio identification: detection of occurrences of a specific audio content (music, advertisements, jingles) in a radio stream.

- Audio motif discovery: detection of repeating objects in audio streams.

- Speaker diarization: segmentation of an input audio stream into homogenous regions according to speaker's identities in order to answer the question "Who spoke when?"

- Nonlinguistic vocalization detection: detection of nonlinguistic sounds such as laughter, sighs, cough, or hesitation.

The paper is organized as follows. In Section 1, each module of the proposed ALISP based compressed representation for generic audio indexing is presented. The radio broadcast database and experimental setup are described in Section 3. The experimental results are presented in Section 4. Conclusions and perspectives are given in Section 5.

1 ALISP-based Data Compression for Generic Audio Indexing

Our goal is to provide a compact representation of audio data on a symbolic level, that can be exploited for further indexing and retrieval. As shown in Figure 1, the proposed system is mainly composed of three modules:

- Automated acquisition (with unsupervised machine learning methods) and Hidden Markov Modeling of ALISP audio models.

- Segmentation (also referred as sequencing and transcription) module that transforms the audio data into a sequence of symbols (using the previously acquired ALISP Hidden Markov Models).

- Comparison and decision module, including approximate matching algorithms inspired form the Basic Local Alignment Search (BLAST) [2] tool widely used in bioinformatics and the Levenshtein distance [15], to search for a sequence of ALISP symbols of unknown audio data in the reference database (related to different audio items).

1.1 Acquisition and Modeling of ALISP Units

In the previous version of ALISP tools [7], the set of HMM models was automatically acquired through parametrization, temporal decomposition, vector quantization, and

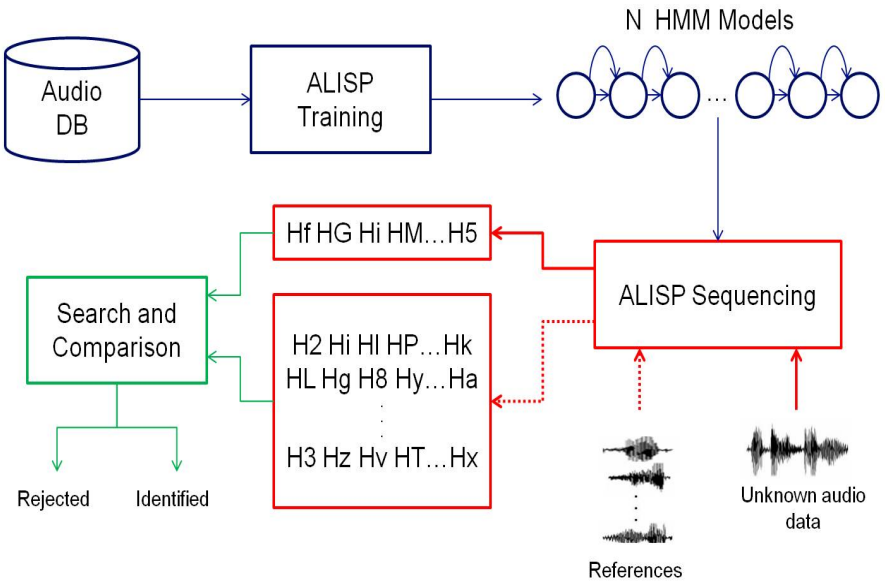

Figure 1: ALISP-based audio indexing system.

Hidden Markov Modeling. In this paper, temporal decomposition is replaced by spectral stability segmentation to speed up the training process of ALISP models.

The parametrization of audio data is done with Mel Frequency Cepstral Coefficients (MFCC). After that, a spectral stability segmentation is computed in order to find the stable regions of the audio signal. These regions represent the spectrally stable segments of audio data. This process is performed using the spectral stability curve obtained by computing the Euclidian distance between two successive feature vectors. The local maxima of this curve represent the segment boundaries while the minima represent the "stable" frames of the audio signal.

The next step is the unsupervised clustering procedure performed via Vector Quantization [16]. This method maps the P-dimensional vector of each segment provided by the spectral stability segmentation into a finite set of M 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 M 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 spectral stability segmentation and Vector Quantization steps. HMM training is performed using the HTK toolkit [1]. 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.

1.2 ALISP Sequencing

Given an observation sequence of features $Y = y_1, ..., y_T$, the recognized ALISP sequence is the one which is the most likely to have generated the observed data Y. In such a way any audio stream is transformed into a sequence of ALISP symbols. An efficient way to solve this problem is to use the Viterbi algorithm [24].

The actual number of ALISP units is 33 (32+silence model) with an average length of 100 ms per model. Compared to the Philips system [11] which extracts 32-bit vector per frame, leading to 5,160 vectors per minute, ALISP methodology provides a very compact way to represent the audio data with approximately 600 ALISP units per minute. Moreover our method is as compact as the audioDNA described in [5] where 800 gens per minute are extracted. This system was only used for audio identification of songs while our system is more generic since it is applied for songs, commercials, speech and laughter.

Therefore, instead of treating the raw audio data directly, ALISP sequencing is performed to create a compressed space in which audio data is represented by a compact sequence of symbols. The audio information retrieval problem is transformed into an approximate string matching problem.

Figure 2 [1] shows a spectrogram of excerpts (2 seconds) from a reference advertisement and two spectrograms of the same advertisement streamed on two different radios with their ALISP transcriptions. Note the presence of some differences between ALISP transcriptions of the three advertisements. These differences could be explained by the similarity between some ALISP classes which leads to confusion during the recognition of these classes.

1.3 Similarity Measure and Searching Method

An important part of the proposed audio indexing system is the matching process. As the main requirement of the proposed audio indexing system is robustness against several types of signal distortions that could be found in radio streams, the actual ALISP unit sequences extracted from an observed signal will not be fully identical to the reference database. The approximate matching step of ALISP sequences is inspired from the Basic Local Alignment Search Tool (BLAST) [2], widely used in bioinformatics.

The BLAST algorithm can be summarized as follows. It is an algorithm for comparing primary biological sequence information, such as amino-acid sequences of different proteins or the nucleotides of DNA sequences. A BLAST search enables to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold. In order to deal with the motif library search, the BLAST algorithm was adapted as shown in Figure 3.

First, a lookup table (LUT) is created by all possible ALISP sequences of w units

[1] This figure is different from the one presented in [12]. The ALISP transcriptions illustrated in this figure are obtained with multi-Gaussian HMM models while the ones in [12] are obtained with the mono-Gaussian HMM models.

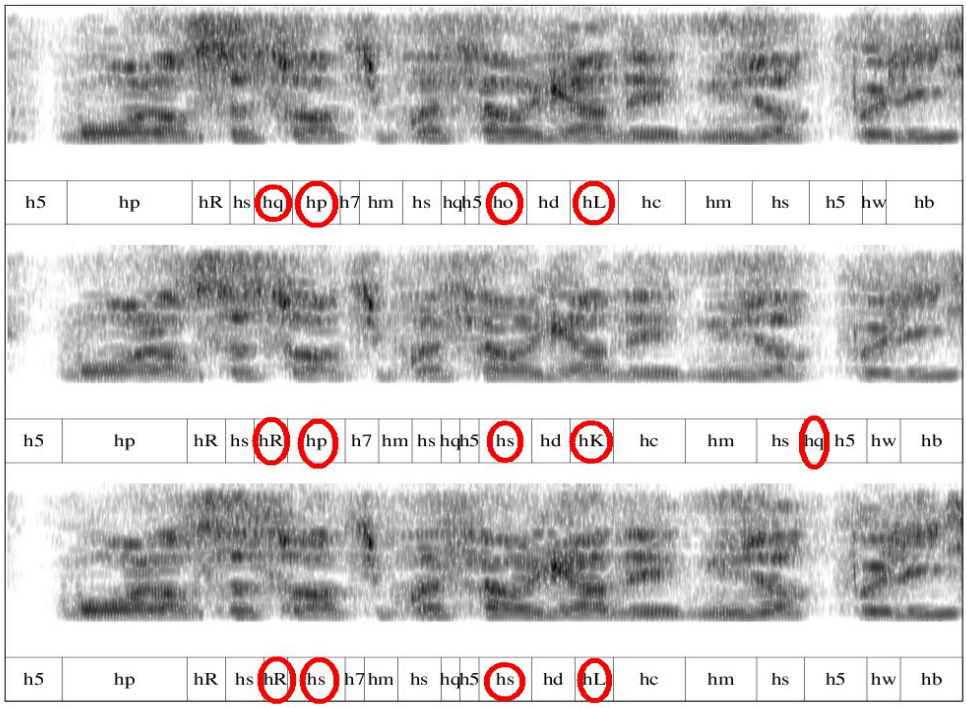

Figure 2: Advertisement spectrograms, taken from the radio broadcast corpus, with their ALISP transcriptions: first spectrogram is an excerpt from the reference advertisement, second one represents the same excerpt from French virgin radio and the last one represents the same excerpt from French NRJ radio

but with an offset of k units that occur in the ALISP transcriptions of the reference database, and each entry in the LUT points to the audio item reference and the position in that item where the respective ALISP unit sequence occurs. Since an ALISP unit sequence can occur at multiple positions in multiple audio items the pointers are stored in a linked list. Thus one ALISP sequence can generate multiple pointers and positions.

Then, during the search phase the ALISP transcription is computed from the query audio stream, and for each subsequence of w units with an offset of k units of that query a set of candidate subsequences is found using the LUT. From this set of subsequences, a list of candidate references and the position where the candidate subsequences occur in that reference is generated for each subsequence of the query data.

The final step of the matching process consists of a simple comparison between the ALISP transcription of the query audio stream and the corresponding candidates using the Levenshtein distance [15]. The candidate audio item selected as the best match is the motif having the lowest Levenshtein distance among all candidates and providing a Levenshtein distance below a certain threshold, determined from a development data.

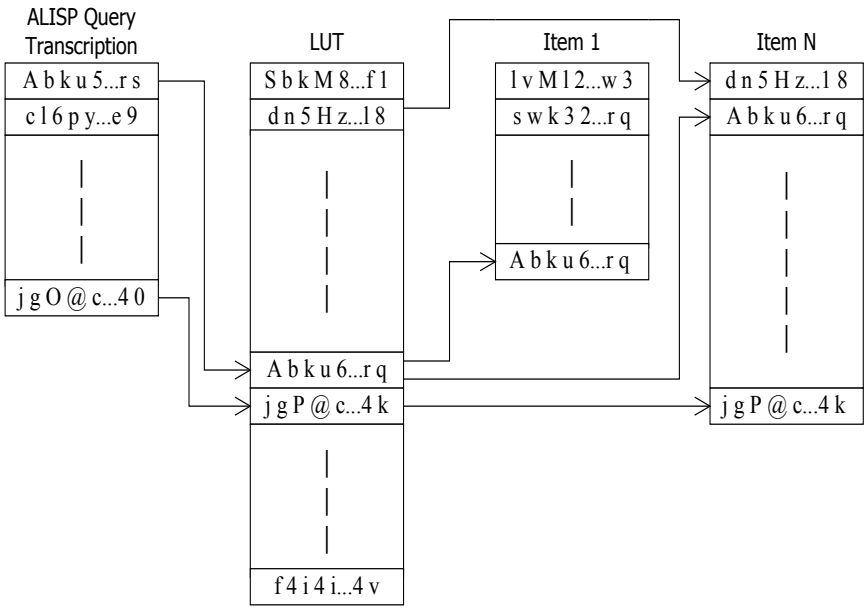

Figure 3: Approximate matching process of an ALISP query transcription using a lookup table (LUT) and a reference database containing N items.

2 Databases

The proposed ALISP-based Data Compression framework for audio indexing is evaluated on four different tasks: audio identification, audio motif discovery, speaker diarization and laughter detection. The audio identification and motif discovery systems are evaluated with a radio broadcast corpus of audio data provided by YACAST (http://www.yacast.fr/fr/index.html). On the other hand, in order to validate our proposal for speaker diarization we participated to the ETAPE'2011 evaluation campaign. Moreover, we use three publicly available corpora to evaluate our system for laughter detection.

Radio Broadcast Corpus: in the framework of the ANR-SurfOnHertz[2] project we had at disposal the YACAST database. This database is split into four subsets:

- Training database: the ALISP HMM models are trained on one day of audio stream from 12 radios (leading to 288 h). It is important to insist that the training database remains the same for all the applications of audio indexing.

- Development database: five days of audio stream are used to study the stability of ALISP transcriptions of advertisements and to set the decision threshold for the Levenshtein distance.

[2]The first author is financially supported by the 2009 ARPEGE program of the French National Research Agency (ANR) under contract number ARPEGE 2009 SEGI17.

- Reference database: it contains 2,172 advertisements and 7,000 songs leading to 9,172 reference items. The advertisement references correspond to the whole commercial item while only a one-minute-long excerpts of each reference song is kept. The position of these signatures within the tracks is unknown. The radio stream from which a given reference was extracted is not part of the evaluation set.

- Evaluation database: seven days of audio stream from three French radios. This data is different from the ones used in the development database and the ALISP training database. This database contains 1,456 advertisements and 4880 songs.

ETAPE Corpus: ETAPE is an evaluation campaign for automatic speech processing [10]. The ETAPE data is divided into three subsets. Train, development and evaluation data (containing respectively 25h30, 8h20, and 8h20).

Laughter Detection Corpora: in order to evaluate the proposed laughter detection system, three publicly available sources are exploited. These databases are SEMAINE-DB [17] (15 hours audiovisual data from 150 participants), AVLaughterCycle [22] (audiovisual laughter database recorded from 24 subjects), and Mahnob laughter databases [20] (3h49min audiovisual laughter data (3h49min) recorded from 22 participants).

3 Results

The evaluation metrics for the audio identification and audio motif discovery systems are precision (P%) and recall (R%) rates.[3]

3.1 Audio Identification Results

R%	P%	Missed Items	False Alarms
93	100	416	0

Table 1: Recall (P%), Precision (R%) values, number of missed items and number of false alarms found in the evaluation set of the YACAST database.

Table 1 shows that the system was not able to detect 416 audio items. These missed items belong to 389 songs and 27 commercials. For music identification, 372 tracks are related to songs that have a different version from the one present in the reference database. For example, 302 live version songs from the test radio stream correspond to the studio version in the references. For commercial identification, the 27 missed advertisements are different from the reference ones. For example, there are 9 commercials spoken by different speakers uttering the same texts. These results show that the proposed system allows to find errors in the manual annotations of songs and advertisements.

[3]Recall : The number of items correctly detected / The number items that should be detected
Precision : The number of items correctly detected / Total number of detected items.

3.2 Audio Motif Discovery Results

Among 4880 songs in the evaluation database, 348 are repeated with a total number of 3081 repetitions. The most repeated motif occurs 24 times while the average number of repetitions is 4. For advertisements, the most repeated motif occurs 16 times while the average number of repetitions is 2. The total number of repeated advertisements is 1315. The results of the experiments are summarized in table 2.

	Rep	R%	P%	MD	FA
Songs	3081	99	100	21	0
Ads	1315	98	99	14	6

Table 2: Number of repetitions (*Rep), precision (P%) and recall (R%) rates, number of missed detection (MD), and number of false alarms (FA), found in the evaluation set of the YACAST database.*

Our results on songs show that the system is not able to detect 21 repetitions. These repetitions are related to songs overlapped with speech which disturbs the detection process. On the other hand, the absence of false alarms shows the ALISP sequencing could easily discriminate the different audio contents. For the advertisements, the system is not able to detect 14 repetitions and leads to 6 false alarms. In fact, these errors are related to the detection of two repetitions of two successive advertisements and one repetition of three successive advertisements. In the manual annotation these repeated advertisements are annotated as separate motifs. This is the origin of these errors.

3.3 Speaker Diarization Results

The proposed system is based on automatically acquired segmental units provided by ALISP tools to search for recurrent segments in TV and radio shows. The reference database is built from audio segments provided by annotated training and development databases. These segments represent speech sentences, silence, noise, jingles, music and advertisements. Then ALISP transcriptions of reference segments are computed using HMMs (Hidden Markov Models) provided by the ALISP tools and compared to the transcriptions of the TV and radio shows stream using the approximate ALISP symbols matching algorithm.

In order to evaluate the speaker diarization system, the Diarization Error Rate (DER) is used. The DER is the sum of three errors: missed detection rate, false alarm rate and speaker error rate. The DER obtained by the proposed system in the ETAPE evaluation campaign is 16.23%. This is the best result in the evaluation campaign among 7 participants, where the greatest DER value is 29.32% [13].

3.4 Laughter Detection Results

Given the high variability of intra and inter speakers for this category of sounds, we decided to use a different approach to search for these sounds from the one used in the previous systems. Our method first adapts ALISP models, previously trained on 288

hours of radio broadcast, using Maximum Likelihood Linear Regression-MLLR [14] and Maximum A Posterior-MAP [9] techniques. The resulting adapted models can then be used to detect local regions of nonlinguistic vocalizations, using the standard Viterbi algorithm.

The proposed system is evaluated on 14 minutes of laughter and 20 minutes seconds of nonlaughter data. The F-measure obtained for the proposed system is 94% while for state of the art systems based on GMMs and HMMs the F-measure are respectively 74% and 88% [18].

3.5 Runtime

Related to the processing time, the computational complexity of the system is mainly limited to the search for the closest ALISP sequence through the Levenshtein distance. The system needs 0.49 seconds to treat one second of signal using the 33 ALISP models on a 3.00GHz Intel Core 2 Duo 4GB RAM, while for the brute search, the system needs 6 seconds to treat one second of signal. It's important to note that the approximate matching technique algorithm speeds up the ALISP transcriptions search without affecting the identification scores.

Conclusions

In this paper we have shown that ALISP-based data compression method provides a good compact representation of audio data, that can be exploited for efficient indexing and retrieval of various types of audio streams. The main idea behind this system is to train a data-driven HMM models that are exploited to transform the raw audio data into a sequence of symbols. Then, the retrieval process is performed using an approximate matching algorithm to compare the ALISP sequences. The proposed system shows good results for audio identification, audio motif discovery, speaker diarization, and laughter detection tasks.

References

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

[2] S. F. Altschul, W. Gish, and W. Miller. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, October 1990.

[3] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370, 2012.

[4] C.J.C. Burges, D. Plastina, J.C. Platt, E. Renshaw, and H.S. Malvar. Using audio fingerprinting for duplicate detection and thumbnail generation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 9–12, 2005.

[5] P. Cano, E. Battla, H. Mayer, and H. Neuschmied. Robust sound modeling for song detection in broadcast audio. *Audio engineering society*, 2002.

[6] J. Černocký. *Speech Processing Using Automatically Derived Segmental Units: Applications to Very Low Rate Coding and Speaker Verification.* PhD thesis, Universit Paris XI Orsay, 1998.

[7] G. Chollet, J. Černocký, A. Constantinescu, S. Deligne, and F. Bimbot. *Towards ALISP: a proposal for Automatic Language Independent Speech Processing*, pages 375–388. 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] J. Gauvain and C.H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *Transactions on Speech and Audio Processing*, 2(2):291–298, 1994.

[10] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, and O. Galibert. The etape corpus for the evaluation of speech-based tv content processing in the french language. In *International Conference on Language Resources, Evaluation and Corpora*, 2012.

[11] J. Haitsma and T. Kalker. A highly robust audio fingerprinting system. In *International Society for Music Information Retrieval*, pages 107–115, 2002.

[12] H. Khemiri, G. Chollet, and D. Petrovska-Delacrétaz. Automatic detection of known advertisements in radio broadcast with data-driven alisp transcriptions. *Multimedia Tools and Applications*, 62(1):35–49, 2013.

[13] H. Khemiri, D. Petrovska-Delacrétaz, and G. Chollet. Speaker diarization using data-driven audio sequencing. In *International Conference on Acoustics, Speech, and Signal Processing*, 2013.

[14] C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 9(2):171–185, 1995.

[15] V. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and control theory*, 10:707–710, 1966.

[16] Y. Linde, A. Buzo, and R.M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, 1980.

[17] 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. *Transactions on Affective Computing*, 3(1):5–17, 2012.

[18] S. Pammi, H. Khemiri, D. Petrovska-Delacrétaz, and G. Chollet. Detection of nonlinguistic vocalizations using alisp sequencing. In *International Conference on Acoustics, Speech, and Signal Processing*, 2013.

[19] P. Perrot, G. Aversano, Raphael Blouet, M. Charbit, and G. Chollet. Voice forgery using alisp: Indexation in a client memory. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 17–20, 2005.

[20] S. Petridis, B. Martinez, and M. Pantic. The mahnob laughter database. *Image Vision Comput.*, 31(2):186–202, 2013.

[21] D. Petrovska-Delacrétaz, C. Černocký, J. Hennebert, and G. Chollet. Segmental approaches for automatic speaker verification. *Digital Signal Processing*, 10(13):198–212, 2000.

[22] J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner. The avlaughtercycle database. In *International Conference on Language Resources and Evaluation (LREC'10)*, pages 2996–3001, 2010.

[23] A. Wang. The shazam music recognition service. *Communications of the ACM*, 49(5):44–48, 2006.

[24] S. Young, N.H. Russell, and J.H.S Thornton. Token passing: a conceptual model for connected speech recognition systems. Technical report, Cambridge University, 1989.