ITRW on Non-Linear Speech Processing (NOLISP 03)
Le Croisic, France
May 20-23, 2003

# Data-driven Speech Segmentation for Language Identification and Speaker Verification

Dijana Petrovska-Delacrétaz[1], Marcos Abalo[1], Asmaa El Hannani[1], and Gérard Chollet[2]

[1] DIVA group, University of Fribourg, Informatics Dept., Fribourg, Switzerland, {dijana.petrovski,marcos.abalo,asmaa.elhannani}@unifr.ch;
[2] ENST, CNRS-LTCI, Paris, France, gerard.chollet@enst.fr

**Abstract.** The common denominator of many speech processing methods is the set of acoustic units chosen to represent the structure of the data. The majority of current systems use phones (or related units) as an atomic representation of speech. The major problems that arise when phone based systems are being developed is the possible mismatch with the data being used and the lack of transcribed databases. The set of speech units can also be learned from examples, like in data-driven approaches. We have already used data-driven acoustic speech units, denoted as Automatic Language Independent Speech Processing (ALISP) units for segmental speaker verification experiments, based on Multiple Layer Perceptrons, and on Dynamic Time Warping (DTW). In this paper we give an overview of the DTW based speaker verification and we present further developments of the data-driven ALISP speech segmentation for language identification experiments. The results confirm the applicability of the proposed method for these two tasks.

## 1 Introduction

The majority of current speech processing systems use phones (or related units) as an atomic representation of speech. Using phonetic speech units leads to efficient representation and implementation for a lot of speech processing systems. The major problems that arise when phone based systems are being developed is the possible mismatch with the data being used and the lack of transcribed databases. For each new language and application, new transcribed databases are needed. This is a serious bottleneck for developing new applications and adapting the existing ones to new languages and tasks. In order to avoid this tedious, error-prone and expensive step of transcribing speech, we need to develop systems that minimize the usage of transcribed databases.

The set of speech units can also be learned from examples, like in data-driven approaches. In [1] a new architecture for speech processing based on units acquired during a data-driven segmentation, that is not grounded on transcribed databases was proposed. These units are denoted as Automatic Language Independent Speech Processing (ALISP) units. In [2], [3] and [4] we have applied this method for speaker verification. In this paper the ALISP units are used for automated language acquisition without transcription. Their utility is evaluated for the language identification task.

There are several reports in the literature on methods for learning spoken language without transcriptions. In [5] 'words' and 'grammar' are automatically acquired from collapsed text. However, that work did not address the issues arising from non-perfect recognition of speech. In [6], it was shown how to acquire lexical units from speech alone without transcriptions and exploit them for spoken language understanding. Several recent experiments [7], [8], [18], report attempts to acquire units from un-transcribed speech using sub-word methods.

All of the above efforts involve learning from speech alone. While one can learn much about a spoken language by merely listening to it, one can progress further and faster by exploiting its associations with a complex environment. This has been demonstrated in both engineering domain [10], and in analysis of children language acquisition [20]. In [12] and[13] we have reported on methods for automatically acquiring salient phone-phrases exploiting semantics. Their utility was evaluated for call-type classification in the "How May I help You" task.

In this paper we report about the application of the ALISP data-driven speech segmentation methods for automatically acquiring salient ALISP sequences for language identification experiments. Language identification systems, currently rely on the informations gathered from the outputs of the automatic speech recognizers of the implied languages ([9], [17], [18], and [19]). The likelihoods of the recognizer outputs are used to decide on the observed language. The speech units involved are phones. In the proposed method, the first step is the automatic acquisition of a common set of acoustic segments from the training data of the languages to be identified, based on ALISP tools. The ALISP sequences could be compared to the multi-grams, as defined in [8]. The next step consists of the search of the language specifics variable length ALISP sequences that characterize the languages to be identified. The final step is the definition of a classification strategy used for the language identification. The utility of the proposed method is evaluated for language identification, on subsets of Swiss French and Swiss German Polyphone databases.

In this paper, we also give an overview of our recent developments of the DTW based speaker verification method using ALISP units. Current best performing text-independent speaker verification systems are based on Gaussian Mixture Models (GMM) [14]. Speech is composed of different sounds and speakers differ in their pronunciation of these sounds. GMM could be interpreted as a "soft" representation of the various acoustic classes that make up the speakers' sounds. They do not take into account the temporal ordering of the feature vectors. The speaker verification approach described in this work is based on speech recognition, grounded on data-driven techniques that do not require neither phonetic nor orthographic transcriptions of the speech data. The main advantages of introducing a speech recognition stage in speaker verification experiments are: to exploit the different speaker discriminant power of speech sounds [15], [2] and to benefit of some higher-level informations resulting from the segmentation. Our speaker verification approach is the following: the segments found in the enrollment speech constitute the speaker specific speech memory, named also *Client-Dictionary*. Another set of speakers is used to represent the non-speaker (world) speech memory, named also as *World-Dictionary*. The speaker verification step is done combining these two dictionaries. The non-linear Dynamic Time Warping (DTW) distance measure is applied in order to find the similarities of an incoming test segment to the client and world dictionaries. The final client score is related to the DTW distance measures.

The outline of this paper is the following: In Section 2 an overview of our recent developments of the DTW based speaker verification method using ALISP units is given. Some examples of the ALISP units are also given. Section 3 describes the proposed language identification method, using the same data-driven speech segmentation methods. Examples of automatically acquired words characterizing the languages to be identified are also given in this section. The conclusions and perspectives are given in Section 4.

## 2 ALISP Based Speaker Verification Experiments

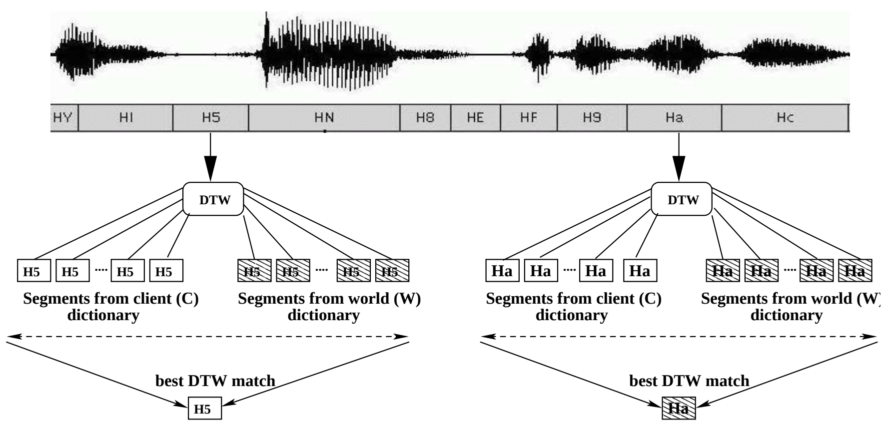


**Fig. 1.** Illustration of the proposed speaker verification method based DTW distance measre.

The steps needed to acquire and model the set of data-driven speech units, denoted as *Automatic Language Independent Speech Processing (ALISP)* units, are described in [4]. The set of symbolic units is automatically acquired through temporal decomposition, vector quantization, segment labeling and Hidden Markov Modeling.

### 2.1 Non-linear Segmental Speaker Modeling

A speaker verification system is composed of training and testing phases. During the training (also known as enrollment phase) the client models are constructed. We propose to use a client's speech memory, build from the ALISP segments, to represent the client enrollment data. The world speech memory, is build with segments found in the speech data representing the world speakers. During the testing phase (see Figure 1), each of the speech segments found is going to be compared with a DTW distance measure, to a set of speech segments representing the claimed speaker identity $S_c$ and the world speakers $S_W$, denoted here as the *Speaker-Dictionary* and *World-Dictionary*. With the DTW distance measure the nearest speech segment from the *Speaker* or *World Dictionary* is found. This is done for all the segments of the test speech data. In the next step, the ratio of the number of times that a segment belonging to the speaker is chosen $nb(S_c)$, versus the mean value of the number of times a world segment is chosen, $mnb(S_W)$, is estimated. This ratio represents the score $S$ of the claimed speaker:

$$score(S) = \frac{nb(S_c)}{mnb(S_W)} .$$

The final decision is taken comparing this score to a threshold (found on an independent development set).

### 2.2 Speaker Verification Database and Results

We have used the National Institute of Standards and Technology (NIST) 2001 cellular speaker verification database, for the development data, used for the *World-ALISP-set*. The results were evaluated on the NIST 2002 cellular data. The *World-ALISP-set* is used for two purposes: to build the gender dependent ALISP recognizers, and to model the world (also called background) speakers. Two gender dependent ALISP speech recognizers are obtained with data coming from 60 female and 59 male speakers. Because in NIST evaluations, the gender is an a priori provided information, we will use gender dependent subsets of each of the above mentioned sets without repeating it explicitly each time. Because of lack of speech data the same speakers are used to model the world speakers.

Linear Prediction Cepstral Coefficients (LPCC), calculated on 20 ms windows, with a 10 ms shift, are used for the acquisition of the initial gender dependent 64 ALISP units. During the Hidden Markov Modeling, we have used the Mel Frequency Cepstral Coefficients (MFCC). They are generally used for common speech and speaker verification purposes. The window and shift values are kept the same as for the LPCC parameterization.

In order to accelerate the search, we have restricted the number of speech units in the dictionary to the 5 longest segments per class and per world speaker, and to the 15 longest ones for the client speaker. For each client speaker the resulting *Client-Dictionary* has about 1'000 segments. The common *World-Dictionary* is represented with 20'000 segments, chosen among the speech data belonging to the world set.

In this paper the proposed speaker verification method is evaluated on the full NIST 2002 cellular data. Without normalization, we obtain an Equal Error Rate (EER) of about 18%. The best results, reported at NIST'2002 evaluations are around 8%. We hope that using further normalization techniques, and exploiting in a better way the speech segmentation, we could improve the presented results.

### 2.3 Examples of Data-driven ALISP Speech Segmentation from Speaker Verification experiments

In order to have an idea of the coherence of the ALISP units, and of their possible correspondence with the phonetic units, we have compared the manually obtained phonetic transcriptions with the automatic segmentation of the same speech data with the ALSIP units.

Figures 2 and 3 show examples of the ALISP segmentation of excerpts from the train and test speech data, spoken by female speakers. For comparison, the manually obtained phonetic segmentation is also given. In Figure 2 there are two occurrences of the English phone sequence *m-ay*. The corresponding ALISP sequences are *HM-Hf-Ha* and *HM-HZ-Ha*. In Figure 3 there is a third occurrence of the phone sequence *m-ay*, recognized as one of the previous ALISP sequences *HM-Hf-Ha*. Such correspondences, have to be validated on transcribed databases.

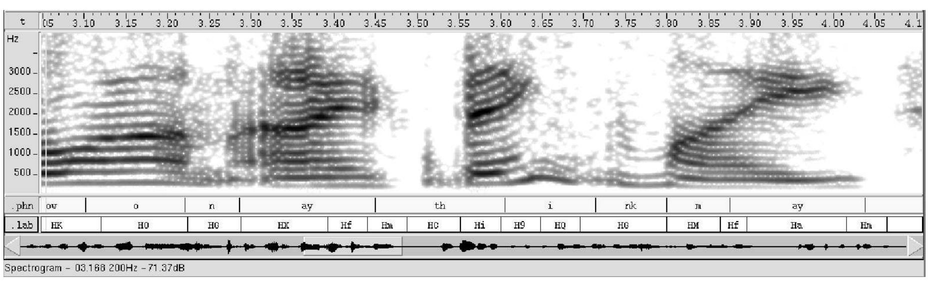


**Fig. 2.** Example of the ALISP segmentation (*.lab* labels) of an excerpt of a train speech data, spoken by a female speaker. For comparison, the manually obtained phonetic segmentation is also given (*.phn* labels).

## 3 ALISP Based Language Identification Experiments

### 3.1 Automatic Acquisition of a Universal Set of ALISP Acoustic Units and of the Language-dependent Sequences

In this paper we report about the application of the ALISP data-driven speech segmentation methods for automatically acquiring salient ALISP sequences for

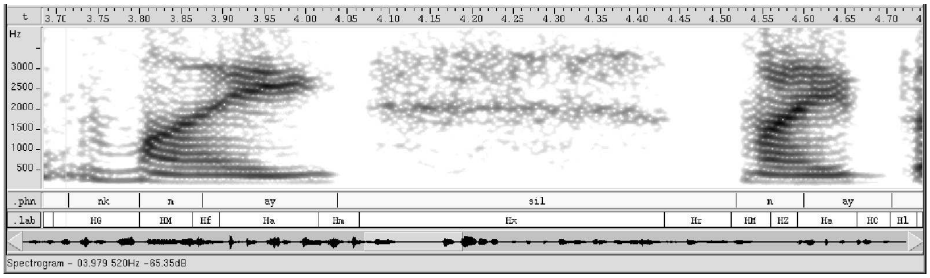


**Fig. 3.** Example of the ALISP segmentation (*.lab* labels) of an excerpt of a test speech data, spoken by a female speaker. For comparison, the manually obtained phonetic segmentation is also given (see *.phn* labels).

language identification experiments. In the proposed method, the first step is the automatic acquisition of a common set of acoustic units from the training data of the languages to be identified. From the training data, we have build a common set of 64 ALISP units characterizing both languages (they are gender dependent), with their corresponding HMM models. We use these HMM models to segment the training and test sentences. The next step is the acquisition of salient ALISP sequences, from the output of the ALISP recognizer using methods of [11]. These ALISP sequences are used for the language identification of the test sentences.

During the test phase, each sentence is first segmented in ALISP units, and among these units exact matches of the language specific ALISP sequences (with a maximum length of 4) are searched. The final step is the definition of a classification strategy used for the language identification on test speech data.

### 3.2 Database for the Language Identification Experiments and Results

The utility of our proposed method is evaluated for language identification, on subsets of Swiss French and Swiss German Polyphone databases[3]. The speech pre-processing parameters are kept identical as for the speaker verification experiments. The number of gender dependent ALISP units common to Swiss German and Swiss French is 64. To have an idea of their classification performance, we have empirically chosen the classification criteria (priority given to longer sequences, and to the maximal number of sequences per language). Among the 5'000 test sentences belonging to female speakers (with a mean length of 3s), in 31% of them no language specific ALSIP-sequences are detected (in the best output of the recognizer). Among the 69% of the test phrases where language specific sequences are detected, using the above mentioned classification criteria,

[3] We thank Swisscom and Robert van Kommer for the availability of these databases.

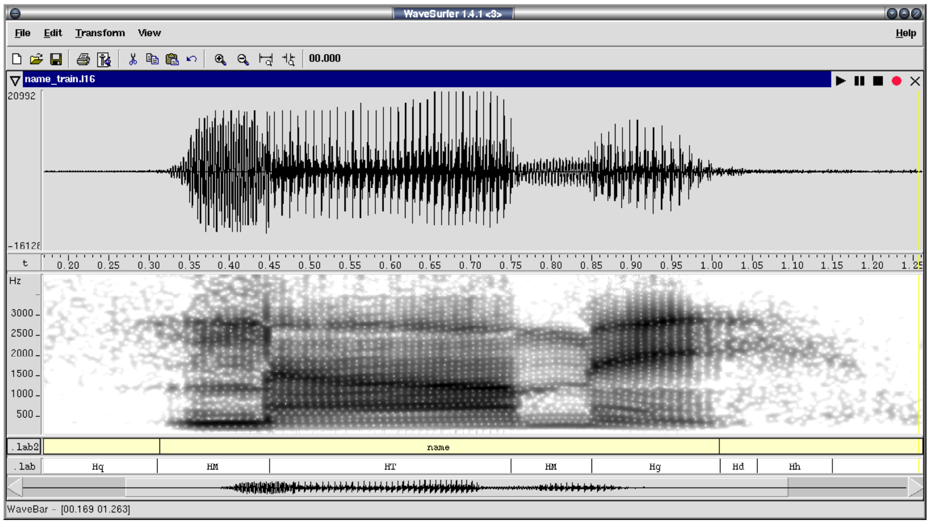


**Fig. 4.** Example of the ALISP segmentation (*.lab* labels) of an excerpt of a train speech data, spoken by a male speaker, corresponding to the German word *name*. For comparison, the manually obtained segmentation is also given (*.phn* labels).

76% of them are correctly classified. These results are slightly worse for the male test data.

An example of the automatically acquired salient french ALISP sequence is *Hp-HH-Ht*. These units correspond to the French word *trois*. Another example is the ALISP sequence *HG-Hp-HT*, that corresponds to the French sub-word *apa*. An example of an automatically acquired ALISP sequence corresponding to German sub-word from the word *name* is the ALISP sequence *HM-HT-HM*. Figure 4 represents the ALISP sequence *HM-HT-HM*, found in the training data. In Figure 5, the same saquence was detected in the test sentence and used to classify it as a German sentence.

We are currently working on the following points: to increase the length of the language specific ALISP-sequences up to 16 units; to search for these sequences not only in the best output of the recognizer, but also in the n-best lattices; to define a better classification method.

## 4 Conclusions and Perspectives

In this paper we have applied a data-driven speech segmentation technique for non-linear segmental speaker verification and language identification. The speaker verification experimental results, are not too far from the current best performing GMM based methods. With the proposed method, no fine tuning of different parameters is done, and more improvements could be foreseen due to

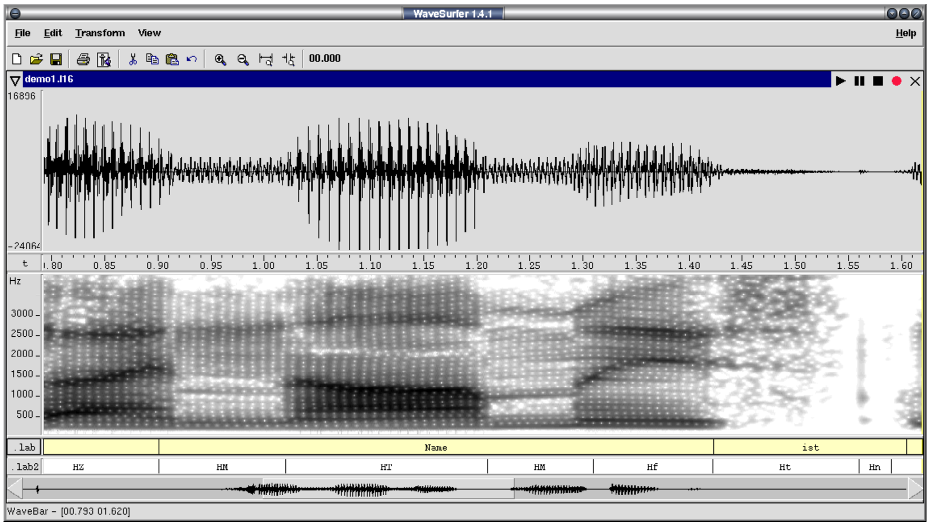


**Fig. 5.** Example of the ALISP segmentation (*.lab* labels) of an excerpt of a test speech data, spoken by a male speaker, corresponding to the German word *name*.

different normalization techniques, or better modeling of the HMM models, or a better choice of the distance measure during the DTW matching.

For the language identification experiments, we are currently working on the following points: to increase the length of the language specific ALISP-sequences up to 16 units; to search for these sequences not only in the best output of the recognizer, but also in the n-best lattices; to define a better classification method.

Data-driven methods are independent of the databases. Therefore, it was straightforward to use the same segmentation method on another database, for the language identification task. In the future we could foresee to combine data-driven methods with speech processing methods based on phonetic units.

## 5 Acknowledgments

Our thanks go to F. Bimbot for his tsd95 package, to J. Černocký for some of the ALISP tools, to Wavesurfer developpers, and to the HTK team.

## References

1. G. Chollet, J. Černocký, A. Constantinescu, S. Deligne, F. Bimbot: Towards ALISP: a proposal for Automatic Language Independent Speech Processing. In Keith Ponting, editor, NATO ASI: Computational models of speech pattern processing Springer Verlag, 1999.
2. D. Petrovska-Delacrétaz, J. Černocký, J. Hennebert, G. Chollet. Text-independent speaker verification using automatically labelled acoustic segments. In International

Conference on Spoken Language Processing (ICLSP), Sydney, Australia, December 1998.

3. D. Petrovska-Delacrétaz D. J. Černocký, G. Chollet. Segmental Approaches for Automatic Speaker Verification. Digital Signal Processing, Special Issue on the NIST'99 evaluations, Vol. 10(1-3), pp. 198–212, January/April/July 2000.
4. D. Petrovska-Delacrétaz, A. el Hannani, G. Chollet, Searching Through a Speech Memory for Text-independent Speaker Verification, AVBPA 2003.
5. D.C. Olivier, Stochastic Grammars and Language Acquisition Mechanism, PhD. Thesis, Harvard University, 1968.
6. A.L. Gorin, S. Levinson,A. Sankar, An Experiment in Spoken Lahguage Acquisition, IEEE Trans. on Speech and Audio, pp. 224-240, vol.2, no.1, part II, Jan. 1994
7. M. Brent, Unsuoervised Learning of Lexical Units. Computational Linguistics 19, pp. 243-262, 1993.
8. S. Deligne, F. Bimbot, Inference of Variable-length Linguistic and Acoustic Units by Multigrams, Speech Communication 23, pp. 223-241, 1997.
9. S. Herbeck, U. Ohler, Multigrams for Language Identgification, Proc. Eurospeech, Budapest, 1999.
10. A.L. Gorin, On Automated Language Acquisition, Journal of the Acoustical Society of America (JASA), 97(6), pp. 3441-3461, June 1995.
11. A.L. Gorin, G. Riccardi, J. Wright, How May I Help You? Speech Communication 23, pp. 113-127, 1997.
12. A.L. Gorin, D. Petrovska-Delacrétaz, G. Riccardi, J.H. Wright, Learning Spoken Language without Transcriptions, Proc. of the ASRU Workshop, Colorado, 1999.
13. D. Petrovska-Delacrétaz, A.L. Gorin, G. Riccardi, J.H. Wright, Detecting Acoustic Morphemes in Lattices for Spoken Language Understanding, Proc. ICSLP 2000, Beijing, China, Dec. 2000.
14. D. A. Reynolds, T. F. Quatieri, R. B. Dunn, Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing, Special Issue on the NIST'99 evaluations, Vol. 10(1-3), pp.19–41, January/April/July 2000.
15. J. P. Eatock, J. S. Mason. A quantitative assessment of the relative speaker discriminant properties of phonemes. Proc. ICASSP, volume 1, pp. 133–136, 1994.
16. J. Olsen. A two-stage procedure for phone based speaker verification. In G. Borgefors J. Bigün, G. Chollet, editor, First International Conference on Audio and Video Based Biometric Person Authentication (AVBPA), pages 199–226, Crans, Switzerland, 1997. Springer Verlag: Lecture Notes in computer Science 1206.
17. J. Navratil, Spoken Language Recognition - A step towards Multilinguality in Speech Processing, IEEE 2001.
18. H.L.Thomas, E.Parris, J.Wright, Reccurent substrings and datafusion for language recognition, ICASSP 2000, vol. 2, pp.169-173.
19. J. Farinas, R.-A. Obrecht, Modelisation phonotactique de grandes classes phonotactiques en vue d'une approche differenciee en identification automatique des langues, Jep 2002.
20. D. Roy, Learning Words from Sights and Sounds: A Computational Model, PhD. Thesis, MIT, 1999.