Automatically Derived Speech Units: Applications to Very Low Rate Coding and Speaker Verification

Jan Černocký[1]*, Geneviève Baudoin[2], Dijana Petrovska-Delacrétaz[3],
Jean Hennebert[3], and Gérard Chollet[4]

[1] Institute of Radioelectronics FEI VUT Brno, cernocky@urel.fee.vutbr.cz
[2] Dpt. Signal et Télécommunications ESIEE Paris, baudoing@esiee.fr
[3] DE–CIRC, EPFL Lausanne, petrovska,hennebert@circhp.epfl.ch
[4] Dpt. Signal et Images, ENST Paris, chollet@sig.enst.fr

Abstract. Current systems for recognition, synthesis, very low bit-rate (VLBR) coding and text-independent speaker verification rely on sub-word units determined using phonetic knowledge. This paper presents an alternative to this approach — determination of speech units using ALISP (Automatic Language Independent Speech Processing) tools. Experimental results for speaker-dependent VLBR coding are reported on two databases: average rate of 120 bps for unit encoding was achieved. In verification, this approach was tested during 1998's NIST-NSA evaluation campaign with a MLP-based scoring system.

1 Introduction

Sub-word units are widely used in various domains of speech processing. Classically, they are based on phonemes or their derivatives (context-dependent phonemes, syllables, etc.) and to be determined, an important amount of phonetic and linguist knowledge is necessary. In order to train a speech processing system, one must dispose of annotated training database (DB). The annotation using phonetically-derived units is a time-consuming, costly and error-prone task. Even if natural language processing can not be done without phonetic and/or linguist expertise, recent advances in Automatic Language Independent Speech Processing (ALISP) [3] have shown, that many tasks relying currently on such knowledge can be performed more efficiently using data-driven approaches. From a practical point of view, this brings revolutionary changes to the methodology of speech processing: extensive human efforts can be replaced by an automated process.

2 ALISP tools

These tools serve for unsupervised search of acoustically coherent speech patterns. They are based on *speech signal data* rather than on the textual representation. The tools are modular and from the ensemble used in coding experiments (Fig. 1), only a sub-set was used in the verification.

* This work is supported by the Ministry of Education, Youth and Sports of the Czech Republic, project No. VS97060

First Workshop on Text Speech and Dialog (TSD'98), Brno, Czech Republic. 1998. pp. 183-188

The *temporal decomposition (TD)* is a representative of algorithms able to detect quasi-stationary parts in the parametric representation of speech. This method, introduced by Atal [1] and refined by Bimbot [2], approximates the trajectories of parameters $x_i(n)$ by a sum of m *targets* a_{ik} weighted by *interpolation functions* (IF):

$$\hat{x}_i(n) = \sum_{k=1}^{m} a_{ik}\phi_k(n), \quad \text{or} \quad \underset{(P \times N)}{\hat{\boldsymbol{X}}} = \underset{(P \times m)}{\boldsymbol{A}} \; \underset{(m \times N)}{\boldsymbol{\Phi}} \tag{1}$$

in matrix notation, where the lower line indicates matrix dimensions. The initial interpolation functions are found using local Singular Value Decomposition with adaptive windowing, followed by post-processing (smoothing, decorrelation and normalization). Target vectors are then computed by: $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{\Phi}^{\#}$, where $\boldsymbol{\Phi}^{\#}$ denotes the pseudo-inverse of IFs matrix. IFs and targets are locally refined in iterations minimizing the distance of $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$. Intersections of interpolation functions permit to define speech segments.

Unsupervised clustering assigns segments to classes. Vector quantization (VQ) is used for automatic determination of classes: class centroids are minimizing the overall distortion on the training set. The VQ codebook $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L\}$ is trained by K-means algorithm with binary splitting. Training is performed using vectors positioned in gravity centers of TD interpolation functions, while the *quantization* takes into account entire segments using cumulated distances between all vectors of a segment and a code-vector. TD with VQ can produce a phone-like segmentation of speech.

Multigrams (MG) [4] may serve for finding *characteristic sequences* of quantized TD events or of segments determined by HMMs. The method is based on finding optimal segmentation of symbol string into *variable length sequences (multigrams)* using likelihood maximization:

$$X^\star = \arg\max_{\forall X} \mathcal{L}(O, X|\{x_i\}), \tag{2}$$

where O is the string of observations, X is the segmentation and $\{x_i\}$ is the codebook of available MGs. The likelihood is given by the product of probabilities $\mathcal{P}(x_i)$ of MGs in the segmentation X. These are not known and must be estimated on the training corpus using iterations of segmentation (2) and of probabilities re-estimation using sequence counts.

Hidden Markov models (HMM) can be used to model the units. HMM parameters are *initialized* using context-free and context-dependent Baum-Welch training with TD+VQ or TD+VQ+MG transcriptions, and *refined* in successive steps of corpus segmentation (using HMMs) and model parameters re-estimation. The speech represented by observation vector string $\boldsymbol{O}$ can then be aligned with models by maximizing the likelihood:

$$\arg\max_{\{M_1^N\}} \mathcal{L}(M_1^N|\boldsymbol{O}), \quad \text{where} \quad \mathcal{L}(M_1^N|\boldsymbol{O}) = \frac{\mathcal{L}(\boldsymbol{O}|M_1^N)\mathcal{L}(M_1^N)}{\mathcal{L}(\boldsymbol{O})}. \tag{3}$$

M_1^N is the sequence of models and $\mathcal{L}(M_1^N)$ is the a-priori probability of M_1^N determined by *language model* (LM).

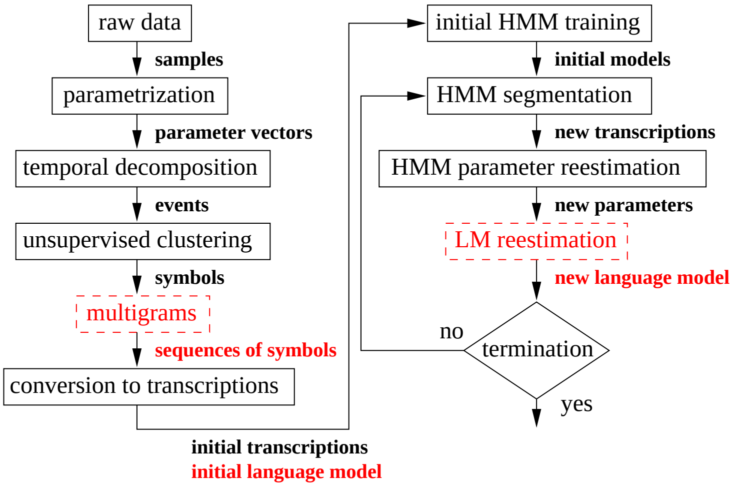

Fig. 1. Data-driven derivation of coding unit set in VLBR phonetic vocoder

3 Very low bit-rate coding

VLBR coding with data-driven units is a framework to test the efficiency and usefulness of the ALISP approach. In this area, the task of pronunciation modelling does not need to be resolved, but the efficiency of algorithms is evaluated by re-synthesizing the speech and by comparing it to the original. If this output is intelligible, one must admit, that this representation is capable of capturing acoustic-phonetic structure of the message and that it is appropriate also in other domains. Moreover (in contrast with classical approach, where the unit set is fixed a-priori and can not be altered), the coding rate in bps and the dictionary size carry information about the *efficiency* of the representation, while the output speech quality is related to its *precision*.

The flow-chart of derivation of coding units (CU) using a training corpus is given in Fig. 1. With these units, the test corpus is encoded (by alignment of HMMs with data) and the efficiency of coding is evaluated by mean bit rate R_u [bps] supposing uniform encoding of sequence indices. Prosody information is not taken into account. *Synthesis* is done using *representatives* drawn from the training corpus. Experimental setup and results are summarized in Tab. 1. In the first case, the synthesis was done by a simple concatenation of representative signals. In BU experiments, the synthesis was LPC with the original prosody. In both sets of experiments, the resulting speech was found intelligible, but the quality is significantly worse than for codecs at several kbps. Details and speech files can be found in [9] and its related Web-page.

4 Text-independent speaker verification

In *text-dependent* experiments, text transcription of the speech sequence used to distinguish the speaker is known. It can serve to align the speech signal into more discriminating classes and an optimized recombination of these class decisions can be done.

Table 1. Summary of VLBR coding experiments

database language speakers	PolyVar Swiss French 1 (the most represented)	BU Radio Speech Corpus American English 2 (F2B, M2B)
parametrization	10 LPCC,ΔLPCC,E,ΔE	16 LPCC,ΔLPCC,E,ΔE
TD	avg. 15 events/sec	avg. 17 events/sec
VQ codebook	64	64
MGs prior to HMMs	no	yes
HMMs to train	1666	64
HMM refinements	1	5
MGs after HMMs	no	yes
coding units	1514	722 (F2B), 972 (M2B)
representatives per CU	8	8
R_u [bps] (test set)	120	110 (F2B), 119 (M2B)

Several studies [5, 6] have demonstrated that some phones show more speaker discriminative power than others, suggesting that a weighting of individual class decisions should be performed when computing the global decision. On the other hand, in *text-independent* systems, the transcription is not available. These systems rely mostly on modelling of the *global* probability distribution function (pdf) of speakers in the acoustic vector space. If one wants to overcome the apparent coarseness of this model and approach a text-independent system to text-dependent one, speech segments of incoming speaker must be aligned with class-dependent models of clients and impostors. This approach is illustrated in Fig. 2. Two possibilities come into account for this alignment: Large Vocabulary Continuous Speech Recognition (LVCSR), which uses previously trained phone models and a language model, or ALISP tools. The later approach was tested experimentally and is described below.

The system was tested on NIST-NSA 1998 data [7]. There are 250 male and 250 female clients, each with more than 2 minutes of training speech. Here, the results are reported only for 30 s test files duration. Data were parametrized using 12 LPCC. The segmentation and classification of speech segments were respectively done by temporal decomposition (avg. of 15 events/sec) and vector quantization with $L = 8$ code-vectors. *Multi-Layer Perceptrons (MLP)* were used to compute client and world probabilities. MLPs worked with a *context* of 4 acoustic frames (2 left, 2 right). For each client, L MLPs, each with 20 hidden nodes, were trained and their scores were summed with equal weights. The performances of segmental verification system were compared with a global one, where only one MLP with 120 hidden nodes was used to distinguish between client and world. The results are reported as DET (Detection Error Tradeoff) curves for two couples of train-test conditions in Fig. 3. We observe comparable but a little worse results of the segmental system when training and test speech come from the same telephone number (SN condition). This ensures us that the segmentation is consistent. As far as more difficult experimental conditions (training and test file from different type of handset – DT) are concerned, the results of the segmental system are better than the global one. Details can be found in [8].

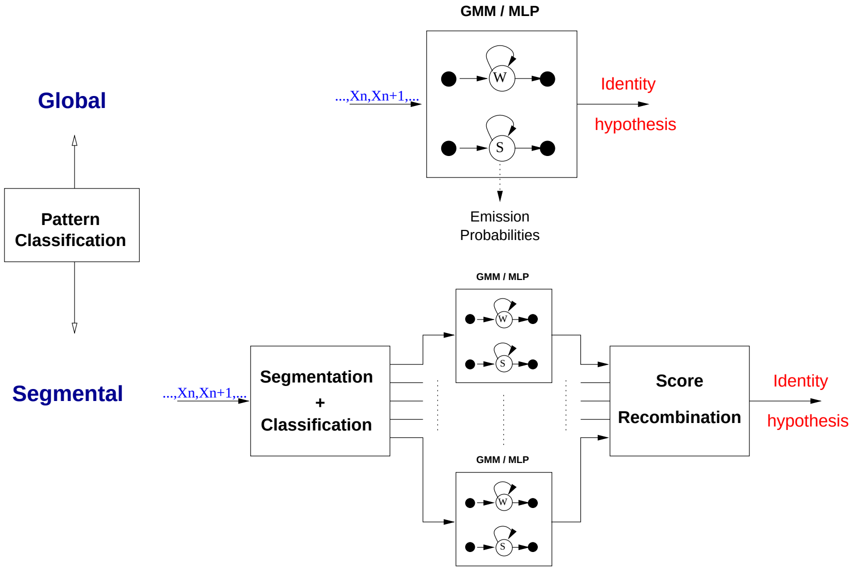

Fig. 2. Global and segmental speaker verification system

5 Discussion and conclusions

In this work, we investigated the use of automatically derived units in speech processing. Their efficiency and usefulness were demonstrated on very low bit-rate coding and text-independent speaker verification. In both domains, the results are promising: in coding, intelligible speech was obtained at mean rate of unit coding $\sim$120 bps and in verification, the segmental system performed better than the global one for a difficult training-test condition. However, a lot of issues are still open, namely the synthesis and speaker/voice adaptation in the former domain and efficient merging of class-dependent scores in the later.

Besides coding and verification, ALISP techniques are potentially useful also in other fields of speech processing: they limit the human interaction necessary (hence the number of errors introduced by humans, and the cost) and they approach the system to the data rather than to units more or less related with the text. However, efforts should be done to apply these methods efficiently in practice.

References

1. B. S. Atal. Efficient coding of LPC parameters by temporal decomposition. In *Proc. IEEE ICASSP 83*, pages 81–84, 1983
2. F. Bimbot. An evaluation of temporal decomposition. Technical report, Acoustic research departement AT&T Bell Labs, 1990
3. G. Chollet, J. Černocký, A. Constantinescu, S. Deligne, and F. Bimbot. *NATO ASI: Computational models of speech pattern processing*, chapter Towards ALISP: a proposal for Automatic Language Independent Speech Processing. Springer Verlag, in press

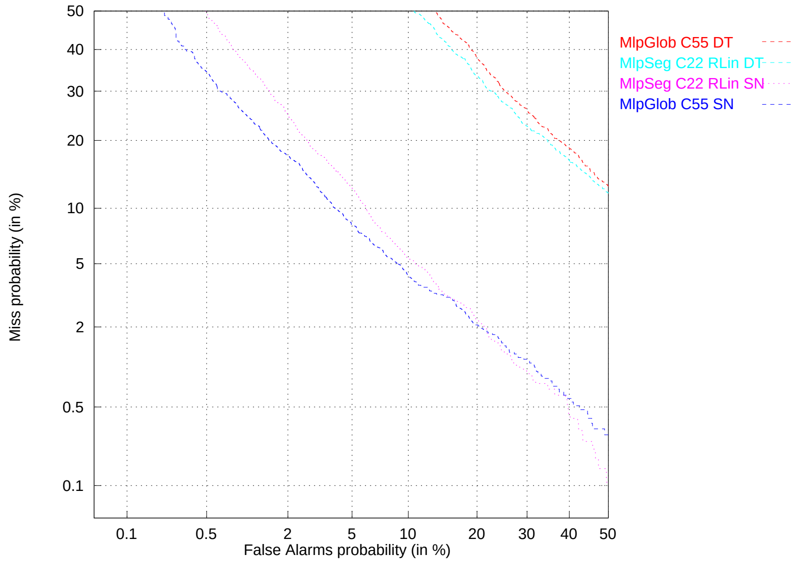

Fig. 3. Global and segmental system, training 2F, test 30 sec, SN and DT conditions

4. S. Deligne. *Modèles de séquences de longueurs variables: Application au traitement du langage écrit et de la parole*. PhD thesis, École nationale supérieure des télécommunications (ENST), Paris, 1996
5. J. P. Eatock and J. S. Mason. A quantitative assessment of the relative speaker discriminant properties of phonemes. In *ICASSP*, volume 1, pages 133–136, 1994
6. J. Hennebert and D. Petrovska-Delacrétaz. Phoneme based text-prompted speaker verification with multi-layer perceptrons. In *Proc. RLA2C 98*, Avignon, France, April 1998
7. A. Martin. The 1998 speaker recognition evaluation plan. Technical report, NIST, March 1998. `http://www.jaguar.nist.gov/evaluations/`
8. D. Petrovska-Delacrétaz, J. Černocký, J. Hennebert, and G. Chollet. Text-independent speaker verification using automatically labelled acoustic segments. In *accepted to International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, December 1998
9. J. Černocký, G. Baudoin, and G. Chollet. Segmental vocoder - going beyond the phonetic approach. In *Proc. IEEE ICASSP 98*, pages 605–608, Seattle, WA, May 1998. `http://www.fee.vutbr.cz/~cernocky/Icassp98.html`