THE USE OF ALISP FOR AUTOMATIC ACOUSTIC–PHONETIC TRANSCRIPTION

Jan Černocký [1], *Geneviève Baudoin* [2] *and Gérard Chollet* [3]

[1] Institute of Radioelectronics, FEI VUT Brno, Czech Republic, cernocky@urel.fee.vutbr.cz
[2] Département Signal et Télécommunications, ESIEE Paris, France, baudoing@esiee.fr
[3] Département Signal et Images, ENST Paris, France, chollet@sig.enst.fr

Abstract: Many current systems for automatic speech processing rely on sub-word units defined using phonetic knowledge. Our paper presents an alternative to this approach – determination of speech units using ALISP (Automatic Language Independent Speech Processing) techniques. Such units were experimentally tested in a very low bit rate phonetic vocoder, where a mean bit rate of 120 bps for unit encoding was achieved. First results of comparison of an ALISP segmentation with a phonetic alignment are presented.

Résumé: Beaucoup de systèmes courants de traitement automatique de la parole sont basés sur des unités de type sous-mot, définies à l'aide du savoir-faire phonétique. Notre contribution présente une alternative à cette approche : une détermination des unités de parole à l'aide des techniques ALISP (Traitement Automatique de Parole, Indépendent de Langue). Nous avons testé expérimentalement de telles unités dans un vocodeur phonétique à très bas débit : le débit moyen ainsi obtenu est de 120 bps. Nous présentons également les premiers résultats de la comparaison d'une segmentation ALISP avec un alignement phonétique.

1. INTRODUCTION

The International Phonetic Association (IPA) sets up as one of its objectives the definition of a symbolic representation of speech for any of the speakers of any language in the world: the International Phonetic Alphabet [1]. However, despite efforts devoted to this topic, some substantial problems persist in the adequacy of this alphabet for spoken speech. Recent advances in ALISP (Automatic Language Independent Speech Processing) [4] led us to the idea of defining such a set of units *automatically*, without an a-priori knowledge; to let it emerge uniquely from the speech data. For this purpose, a number of tools which proved their efficiency in automatic speech processing (coding, recognition, synthesis, language identification, speaker verification) have been developed: temporal decomposition (TD), non-supervized clustering, multigrams (MGS), Hidden Markov Models (HMM) and others. Basic information about these tools with references are given in Section 2.

On contrary to IPA, where it is difficult to find an objective criterion, the set of units can be evaluated using *very low bit rate (VLBR) speech coding* at about 200 bps [9]. At these rates, a symbolic representation of the incoming speech is required. If the decoded speech is intelligible, one must admit that the symbolic representation is capable of capturing the significant acoustic-phonetic structure of the message. Moreover, the coding rate in bps and dictionary size give an idea of *efficiency* of the description while the quality of decoded speech is related to its *precision*. Section 3 gives an overview of our VLBR coding experiments and their results. However, the domain with the greatest need of optimized and automatically derivable units is the large vocabulary continuous speech recognition (LVCSR) based in current systems on phones or their derivatives (context-dependent phones, syllables). Section 4 presents first results of a comparison of two alignments of data (phonetic and ALISP) in terms of a confusion matrix. It also contains some reflections on encoding of target vocabulary using data-driven units.

2. ALISP TOOLS

Data-driven units were searched using ALISP tools presented in Fig. 1. These tools are modular and for example the multigrams can be positioned before or after HMM modelling. Language model (LM) modules determine and

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

refine a-priori probabilities of units — they were used only in previous experiments with PolyVar [9].

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 [2] and refined by Bimbot [3], 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 minimising the distance of $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$. Intersections of interpolation functions permit to define speech segments.

Unsupervized clustering assigns segments to classes. Vector quantization (VQ) is used for automatic determination of classes: class centroids are minimising 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) [5] 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 can then be aligned with models by likelihood maximization.

3. VERY LOW BIT RATE CODING: EXPERIMENTS AND RESULTS

First experiments with VLBR coding were performed with one speaker from the Swiss-French database PolyVar. The results are described in [9]. Here we will concentrate on experiments conducted with speaker F2B from Boston University Radio Speech Corpus [8]. The parametrization was done by 16 LPCC coefficients (completed by ΔLPCC, energy and Δenergy for HMMs). The temporal decomposition[1] produced 17 targets per second in average. The clustering was done with VQ codebook consisting of L=64 code-vectors. No multigrams were used prior to HMM training and the models were trained directly on TD+VQ transcriptions. After initial model training, 5 iteration of refinement (HMM segmentation and parameter re-estimations) were performed. The HMM alignment likelihood increased, indicating an improvement of acoustic match between data and models. The synthesis was done using segments drawn from the training corpus and a LPC synthesizer. Experiments were done also with *post-processing* of last generation HMM segmentation by multigrams. Multigrams with n=6 were trained with resulting numbers N_1=64, N_2=311, N_3=260, N_4=68, N_5=16, N_6=3.

For F2B, the resulting rate for coding of unit indices (assuming uniform encoding) was 126 bps for the last generation HMM segmentation and 110 bps when segments were post-processed by multigrams. The decoded speech was found intelligible, but of much lower quality than for codecs operating at several kbps. Smoothing at segment borders and a better quality synthesis are among the open problems. However, the principle of the approach was confirmed, as passing the speech through a fully automatically trained symbolic representation preserved the information contained in the message. Speech files can be downloaded from the Web-page related to [9].

[1] Thanks to Frédéric Bimbot for the permission to use his package `td95`.

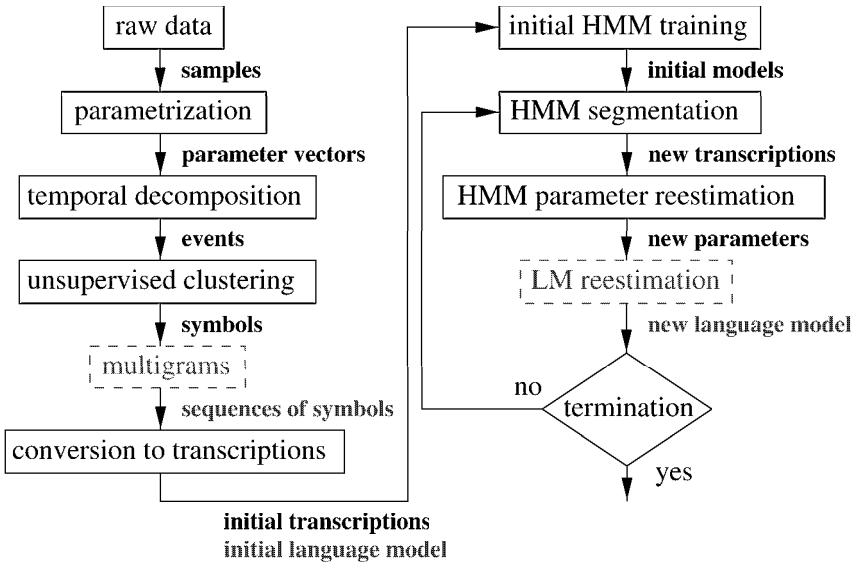

Figure 1: Data-driven derivation of speech unit set in VLBR phonetic vocoder.

4. COMPARISON WITH PHONETIC ALIGNMENTS

The phonetic alignments available with BU corpus allowed us to investigate the correspondence of phones and ALISP units. These alignments were obtained at BU using a segmental HMM recognizer constrained by possible pronunciations of utterances [8]. In our comparison, the alignment files without hand-corrections (extension `.lba`) were used. Phonetic alignments were taken as reference and ALISP segmentations (last generation HMM) were compared against them. The measure of correspondence was the relative overlap r of ALISP unit with a phoneme. The results are summarized in *confusion matrix* $\boldsymbol{X}$ $(n_p \times n_a)$, whose elements are defined:

$$x_{i,j} = \frac{\sum_{k=1}^{c(p_i)} r(p_{i_k}, a_j)}{c(p_i)}. \tag{3}$$

n_p and n_a are respectively the sizes of phoneme and ALISP unit dictionaries, p_i is the i-th phoneme, a_j is the j-th ALISP unit, $c(p_i)$ is the count of p_i in the corpus and $r(p_{i_k}, a_j)$ is the relative overlapping of k-th occurrence of p_i with ALISP unit a_j. The columns of $\boldsymbol{X}$ are rearranged to let the matrix have a quasi-diagonal form[2] and the resulting matrix is given in Fig. 2. On contrary to BU alignments, where stressed vowels are differentiated from unstressed ones, we used the original TIMIT [7] phoneme set.

Although these experiments showed a correlation of phonemes and ALISP units, an ALISP recognition system should not be based on direct phoneme–ALISP mapping. It would be more efficient to represent the target dictionary as probabilistic combinations of *sequences* of ALISP units. The work of Fukada [6] on phoneme- and word-based ASU (automatically derived segment unit) composition, and Deligne's joint multigrams [5] bring interesting insights on this representation.

5. DISCUSSION, CONCLUSIONS

In this paper, we described an alternative approach to phonetically derived speech units — their data-driven ALISP determination. The power of such representation was demonstrated on speaker-dependent phonetic-like VLBR coding, where intelligible speech was obtained at mean rate of 120 bps for unit encoding. Possible links of phonetic alignment with ALISP representation were studied — although those results are preliminary, they show correspondence of some ALISP units with phonemes or phonetic classes. However, rather than direct phoneme–ALISP mapping, the goal is a representation of a speech recognizer's vocabulary in terms of those new units and a comparison of its performances with a "classically" defined scheme.

[2]Thanks to Vladimír Šebesta and Richard Menšík for help with visualization of confusion matrices.

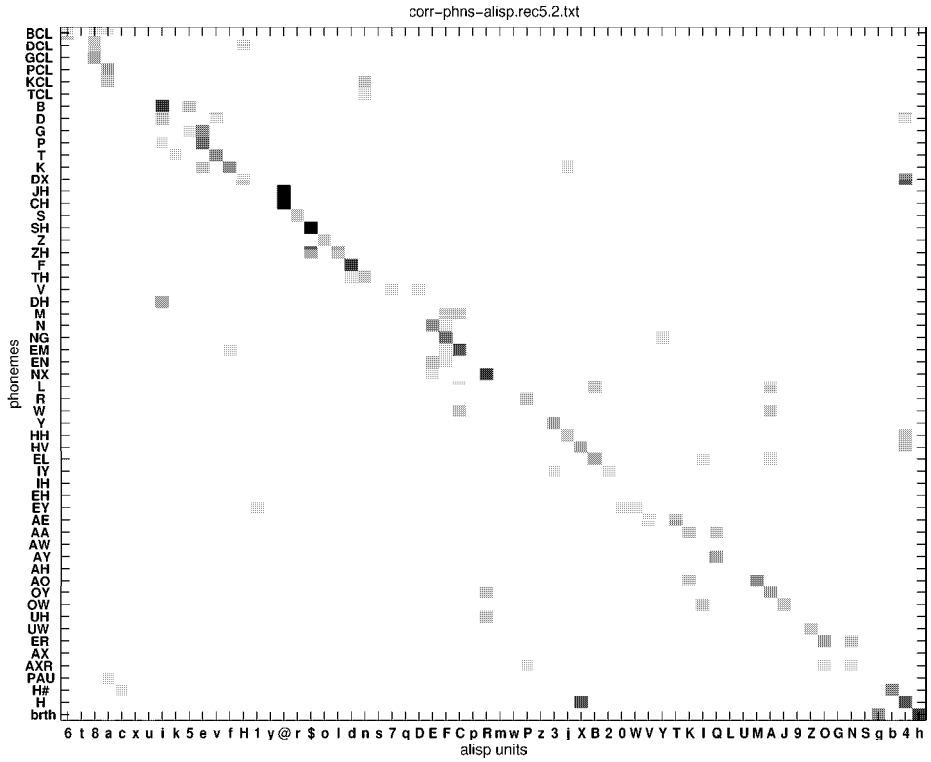

Figure 2: Correspondence of ALISP segmentation and phonetic alignment for speaker F2B in BU corpus. White color corresponds to zero correlation, black to maximum value $x_{i,j}$=0.806

6. REFERENCES

[1] International Phonetic Association (IPA) homepage. http://www.arts.gla.ac.uk/IPA/ipa.html.

[2] B. S. Atal. Efficient coding of LPC parameters by temporal decomposition. In *Proc. IEEE ICASSP 83*, pages 81–84, 1983.

[3] F. Bimbot. An evaluation of temporal decomposition. Technical report, Acoustic research departement AT&T Bell Labs, 1990.

[4] 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.

[5] 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.

[6] T. Fukada, M. Bacchiani, and K. Paliwal Y. Sagisaka. Speech recognition based on acoustically derived segment units. In *Proc. ICSLP 96*, pages 1077–1080, 1996.

[7] J.S. Garofolo, L.F.Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N.L. Dahlgren. DARPA–TIMIT acoustic–phonetic speech corpus. Technical Report NISTIR 4930, U.S. Department of Commerce, National Institute of Standards and Technology, Computer Systems Laboratory, February 1993.

[8] M. Ostendorf, P.J. Price, and S. Shattuck-Hufnagel. The Boston University radio news corpus. Technical report, Boston University, February 1995.

[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.