INFERENCE OF VARIABLE-LENGTH ACOUSTIC UNITS FOR CONTINUOUS SPEECH RECOGNITION

*Sabine DELIGNE** *Frédéric BIMBOT*

Télécom Paris / ENST - Dept Signal, CNRS - URA 820, 46 Rue Barrault,
75634 Paris cedex 13, France, European Union.
e-mail : deligne@sig.enst.fr, bimbot@sig.enst.fr

ABSTRACT

In the field of speech recognition, the patterns assumed to structure the speech material (phonemes, triphones, words...) are defined *a priori* according to a linguistic criterion, whereas the recognition criterion is based on an acoustic similarity measure. From this may result a lack of consistency for the recognition units. In this paper, we explore the possibility of a more data-driven approach, where recognition units are derived according to an acoustic criterion, and then, mapped to variable length sequences of phonemes in an unsupervised way. Continuous speech recognition experiments are reported to evaluate the consistency of those units as opposed to linguistically defined units.

1. MOTIVATIONS

Continuous speech recognition performance is affected by the choice of the appropriate unit for acoustic modeling. As an alternative to the traditionnal linguistically defined units, some effort is being put into the definition of recognition units based on an acoustic criterion, like for instance in [1] [2]. In those works, the phonetic content of the recognition units is not known *a priori*, but usually the number of distinct units, and the length of the associated models are set in advance. Conversely, in this paper, we propose a procedure for deriving acoustic units, the number and the length of which also result from the derivation process. In section 2., we present the general multigram framework on which our approach relies, and in section 3., we show how it can be applied to infer acoustic recognition units. Finally, in section 4., we report experiments on continuous speech recognition, where the performances of systems based on either acoustic multigrams or linguistic units are compared.

2. THE MULTIGRAM FRAMEWORK

2.1. Formulation of the Multigram Model

The multigram approach [3] [4] can be understood as a production model. A source emits a string $\mathbf{Z}$ of units, called multigrams, drawn from a limited set $\{\mathbf{z_i}\}$. Each multigram gives rise to a variable-length sequence of elementary observations. The only observable output of this process is the string of observations $\mathbf{O}$, resulting from the concatenation of all sequences. Given an observed string $\mathbf{O}$, we want to retrieve the set of distinct underlying multigrams $\{\mathbf{z_i}\}$, and to identify in $\mathbf{O}$ the observation sequences originating from a common multigram. Locating multigrams in a string $\mathbf{O}$ involves finding a segmentation $\mathbf{S}$ of $\mathbf{O}$:

$$\begin{array}{ll} \mathbf{Z}: & \mathbf{z_{(i_1)}} \qquad\qquad \mathbf{z_{(i_2)}} \qquad\qquad \mathbf{z_{(i_3)}} \quad \cdots \\ & \Uparrow \qquad\qquad\quad \Uparrow \qquad\qquad\quad \Uparrow \\ \mathbf{S}: & [\ \mathbf{o_{(1)}}\ \mathbf{o_{(2)}}\ ] \oplus \quad [\ \mathbf{o_{(3)}}\ ] \oplus \quad [\ \mathbf{o_{(4)}}\ \mathbf{o_{(5)}}\ \mathbf{o_{(6)}}\ ] \quad \cdots \\ & \qquad\qquad\quad \Uparrow \\ \mathbf{O}: & \mathbf{o_{(1)}}\ \mathbf{o_{(2)}}\ \mathbf{o_{(3)}}\ \mathbf{o_{(4)}}\ \mathbf{o_{(5)}}\ \mathbf{o_{(6)}}\ \cdots \end{array}$$

The optimal set of multigrams is derived by maximizing jointly both the likelihood of the data and of the set $\{z_i\}$:

$$\{z_i\}^* \ = \ \arg\max_{\{z_i\}} \ \mathcal{L}(O \mid \{z_i\})\ \mathcal{L}(\{z_i\}) \qquad (1)$$

The first term in (1), $\mathcal{L}(O \mid \{z_i\})$, measures how well the data fits a given set $\{z_i\}$. It is computed as:

$$\mathcal{L}(O \mid \{z_i\}) \ = \ \sum_{(S,Z)} \mathcal{L}(O,\ S,\ Z \mid \{z_i\})$$

The second term in (1), evaluates the likelihood of the set $\{z_i\}$ itself[1]. The *a priori* distribution of all possible sets is not known, but, according to information theory, $\mathcal{L}(\{z_i\})$ is related to the number of bits required to fully specify the set $\{z_i\}$. Including this term in the optimization criterion aims at balancing the best fit to the data with the least complexity, evaluated as a number of bits. It is based on MDL[2] approaches [5], the expected advantage of which is to reduce the risk of overlearning frequently noticed whith ML estimations [6]. The segmentation and the string of patterns assumed to underly O are:

$$(S^*,\ Z^*) \ = \ \arg\max_{(S,Z)} \ \mathcal{L}(O,\ S,\ Z \mid \{z_i\}) \qquad (2)$$

2.2. Case of Independent Multigrams

In the case multigrams are assumed to be independent, the likelihood of data structured by $(S,\ Z)$ can be expressed as:

$$\mathcal{L}(O,\ S,\ Z \mid \{z_i\}) \ = \ \prod_t \ p(s_{(t)} \mid z_{i_t})\ p(z_{i_t}) \qquad (3)$$

where $s_{(t)}$ denotes the observation sequence corresponding to the multigram z_{i_t}, of rank t in Z. The model is thus fully described by the prior distribution of the multigrams, $\{p(z_i)\}$, and by the set of distribution functions, $\{p(... \mid z_i)\}$, which characterize the variability of the sequences observed for each multigram z_i. The maximization of the data likelihood expressed in (3) tends to favor the inference of highly recurrent multigrams having a low variability. Besides, as specifying a multigram of probability $p(z_i)$ requires a minimum of $-\log p(z_i)$ bits, the minimization of the complexity leads to disqualify multigrams of low probability.

*We are grateful to Stéphanie Dubost and Corinne Grison-Nave for the development of the reference systems used in section 4..

[1] In traditionnal approaches, this set is usually fixed to a predefined set of linguistic units, and its likelihood value is not taken into account.

[2] Minimum Description Length

3. APPLICATION TO SPEECH RECOGNITION

3.1. Reformulation of the Recognition Process

Recurrent patterns extracted from strings of acoustic observations can be used as acoustic units for speech recognition. A recognition task is usually formulated as the determination of the linguistic string L^*_{te}, of maximum likelihood, given an acoustic string O_{te}:

$$L^*_{te} \;=\; \arg\max_{L}\; \mathcal{L}(L \mid O_{te}) \;=\; \arg\max_{L}\; \mathcal{L}(O_{te},\; L)$$

The set $\{z_i\}$ defines a level of intermediate representation Z between the acoustic and linguistic levels, so that:

$$L^*_{te} \;=\; \arg\max_{L} \sum_{Z}\; \mathcal{L}(O_{te},\; Z,\; L) \qquad (4)$$

To reduce computational costs, the assumption is made that there exists a single optimal intermediate representation Z^*_{te}, which accounts for most of the data likelihood, and Z^*_{te} is searched for independently from L. The decoding of a test string O_{te} is thus a 2-step procedure: first, an acoustic decoding, to retrieve the most likely underlying structure $(S^*_{te},\; Z^*_{te})$, and second, a linguistic decoding, to retrieve the phonetic string $\widehat{L}^*_{te}$, which most likely matches Z^*_{te}:

$$\widehat{L}^*_{te} \;=\; \arg\max_{L}\; \mathcal{L}(Z^*_{te},\; L) \qquad (5)$$

To compute $\mathcal{L}(Z^*_{te},\; L)$, a probabilistic model enabling to measure the quality of a mapping between an intermediate representation and a linguistic representation has to be preliminary estimated. At this stage of the decoding, a linguistic component can be used to evaluate the likelihood of L, just as is usually done in a classical recognition system:

$$\widetilde{\widehat{L}}^{*}_{te} \;=\; \arg\max_{L}\; \mathcal{L}(Z^*_{te} \mid L)\; \mathcal{L}(L) \qquad (6)$$

3.2. Inference of the Acoustic Units

• **Nature of the acoustic observations** The observed string O is a stream of continuous-valued vectors, issuing from Temporal Decomposition (TD). TD [7] is a model of spectral evolution, which describes a speech segment as a linear combination of a limited set of vectors called targets. The temporal contribution of each target is expressed by an interpolation function (cf. Figure 1). As spectral characterizations of acoustic events, the target vectors are expected to show less variability than the original frames, hence multigrams are searched for as variable-length sequences of TD target vectors.

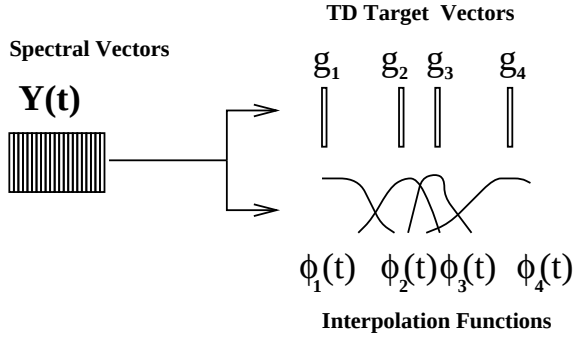

Figure 1.
A graphic illustration of Temporal Decomposition (TD)

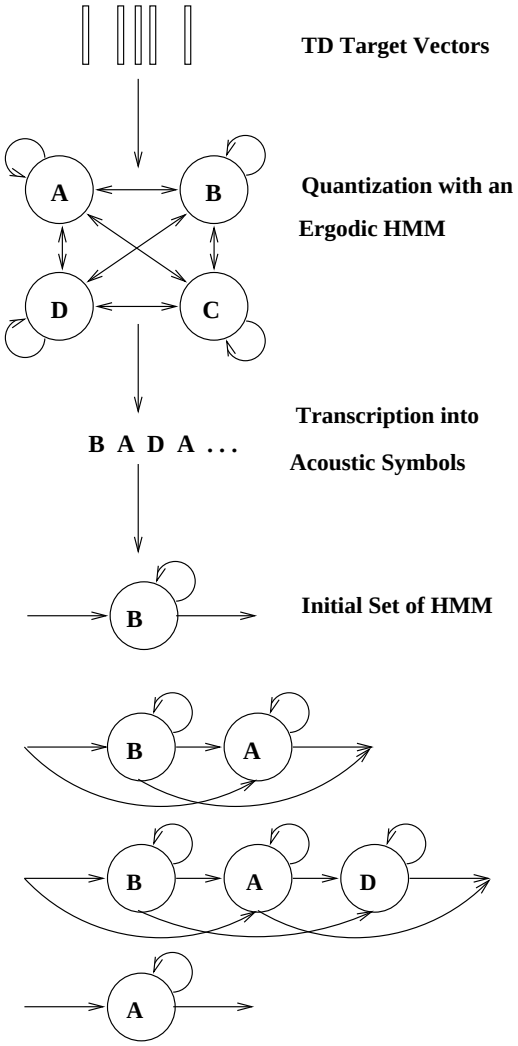

Figure 2. *Definition of an initial set of HMM.*

• **Characterization of each multigram** Each multigram z_i is associated to a Hidden Markov Model (HMM), characterizing its modeling variability. Consequently, the inference of a set of multigrams can be viewed as the inference of a set of HMM. Note that it is altogether the number of distinct HMM, as well as their number of states, which need to be inferred.

• **Definition of an initial set of HMM (cf. Figure 2)** Each target vector is quantized and replaced by a symbol, called acoustic symbol, denoting its quantization class. In the resulting string of acoustic symbols, any combination of n $(n \leq n_{max})$ symbols occurring more than a prespecified threshold, leads to define a left-to-right n-state HMM. Each state parameters are initialized with the mean and covariance values of the corresponding quantization class.

• **Iterative reduction of the set of HMM** The data likelihood and the complexity of the model are alternately optimized, through an EM procedure. Given $(S^{*(k)},\; Z^{*(k)})$ the most likely underlying structure at iteration (k), iteration $(k+1)$ consists of two steps:

- modification of the set $\{z_i\}^{(k)}$ into $\{z_i\}^{(k+1)}$ to reduce its complexity: multigrams occurring less than a prespecified number of times in $(S^{*(k)},\; Z^{*(k)})$ are removed from $\{z_i\}^{(k)}$,
- reestimation of the parameters of the remaining models to

maximize the data likelihood, $\mathcal{L}(O \mid \{z_i\}^{(k+1)})$; the *a priori* probability of each multigram is reestimated as its relative frequency along $Z^{*(k)}$, and the parameters of its HMM are reestimated using the Baum-Welch algorithm.
Iterations are stopped when the set of units becomes stable. Note that a major drawback of this inference process is that it does not offer the possibility to create new HMM, even though it might improve the accuracy of the acoustic modeling, without dramatically increasing its complexity.
• **A posteriori mapping with phonetic sequences** The HMM issuing from the last iteration of the inference process are used to produce the most likely transcription of the target vectors into acoustic symbols. It is obtained by reporting the acoustic symbols corresponding to the states visited during a Viterbi recognition procedure. The transcription into acoustic symbols is aligned on the transcription into phonetic symbols, through a probabilistic many-to-many mapping based on the joint multigram model [8]. This procedure jointly parses the two strings according to a ML criterion, matching variable-length sequences of acoustic symbols with variable-length sequences of phonemes. It results in a dictionary, where pairs of acoustic and phonetic sequences are assigned a probability of co-occurrence. This mapping is performed regardless of the boundaries of the words in the utterances, which are not known.

3.3. Acoustic and Linguistic Decoding

During the acoustic decoding, the most likely transcription into acoustic symbols of a test stream of TD target vectors is retrieved, using the HMM issuing from training. Then, it is decoded into a string of phonemes (Equation (5) or (6)) using the probabilistic mapping provided by the dictionary of joint multigrams.

4. EVALUATION OF ACOUSTIC MULTIGRAMS

We report comparative evaluations of several recognition systems based on multigrams, and of reference systems[3].

4.1. Preparation of the Database

Recognition experiments are conducted on a French database of continuous speech, consisting in sailing weather forecasts (vocabulary of about 400 words), uttered by a single male speaker, and digitally recorded from the radio at 16 kHz. 60 minutes of speech are used for training, and 30 minutes for testing. TD target vectors are computed from 10 ms frames of 16 LPCC parameters.

4.2. Experimental Protocol

Acoustic component based on multigrams The inference procedure described in section 3.2. is applied, using either 32, 64 or 128 distinct acoustic symbols to quantize the target vectors. A HMM is defined for any combination of at most 5 acoustic symbols, occurring more than 20 times; consequently, HMM may have 1 up to 5 states. During the inference process, HMM having a number of occurrences which is less than 10 are discarded. The process is stopped after 5 iterations, as it was empirically noticed to end up with a roughly stabilized set of HMM. During the joint multigram mapping procedure, sequences of up to 5 acoustic symbols are paired with sequences of up to 5 phonemes.

Acoustic component based on linguistic units Three reference systems based on recognition units which are either phonemes, triphones, or words are build, also using the TD target vectors as training observations. Their acoustic components comprise, respectively:

[3] The estimation and recognition procedures involving HMM are performed using the HTK tool kit.

- 35 HMM modeling the French phonemes,
- 1599 HMM modeling the triphones occurring more than 3 times in the training utterances; the 35 models of the phoneme based system are added to this set to ensure an exhaustive phonetic coverage,
- 401 HMM modeling the words of the training vocabulary.

Besides, a 1-state HMM modeling silence, and a 1-state HMM modeling breath noise are added to each reference system. Due to the use of TD target vectors as state emissions, all HMM in the reference systems have only 1 state per phoneme[4]. As the database is not segmented into phonemes, nor even into words, the initial models are aligned on the training stream of target vectors, proportionnally to their number of states. Then, a step of reestimation of the model parameters and a step of realignment of the models on the training set are alternately repeated, till a converged alignment is reached.

Linguistic Component Apart from the training of the acoustic components of all systems, a bigram language model is estimated, which assigns an uniform probability distribution to all unknown combinations of 2 words. The perplexity values computed with this model on the training and on the test sets are respectively 6.3 and 9.8. For bigrams of words to be usable within the multigram recognition framework (decoding according to (6)), the succession of the decoded phonetic sequences must comply with constraints of lexicality. At each step of the decoding process, the set of possible candidates is restricted to those sequences of phonemes only, the concatenation of which still preserves the lexicality of the resulting string. In the case of a system based on phonemes, or on triphones, the HMM are concatenated to form models of words.

4.3. Analysis of the results

Number of quantization classes The results obtained by using either 32, 64 or 128 distinct acoustic symbols are in Table 1. As the number of quantization classes is higher, the number of distinct acoustic units increases. It is to be interpreted as an over partitionning of the speech sequences, since the average number of acoustic sequences matched with a common phonetic sequence, after the joint multigram mapping process, also increases. The recognition scores deteriorate accordingly from 76.1 % to 69.1 % word accuracy for experiments based respectively on 32 and on 128 acoustic symbols. Thus, it seems that, on a single speaker database with a limited vocabulary, a clustering into a relatively small number of classes (32) is enough.

Temporal decomposition Some experiments were conducted on the initial frames to evaluate how suitable a spectral representation based on TD is for the inference of acoustic units. The quantization of the frames, instead of the TD targets, produced a highly instable succession of acoustic symbols (alterned repetitions of 2 symbols for instance), so that it was not used to initialize a set of HMM. But, once an initial set of HMM is defined using a string of quantized TD targets, the inference process can be pursued with the initial frames as state emissions[5]. In our experiments, it results into an increased number of insertions: when using 64 distinct acoustic symbols, the phonetic accuracy on the test set deteriorates from 68.6 % to 64.4 % (while the percentage of phonemes correctly identified remains unchanged).

[4] The number of TD target vectors is 4.7 less than the number of initial frames, so that the average number of target vectors per phoneme approximately equals 1.6.

[5] The temporal information necessary to locate the frames corresponding to each TD target is provided by the interpolation functions.

Number of distinct acoustic symbols		
32	64	128
Number of HMM		
613	755	875
Average number of acoustic sequences per sequence of phonemes		
1.7	1.9	2.2
Phonetic accuracy with no language model		
70.6	68.6	66.2
Word accuracy with a bigram model		
76.1	74.2	69.1

Table 1.
Comparison of systems based on multigrams.

Acoustic modeling of			
phonemes	triphones	words	multigrams (32)
Number of HMM			
37	1636	403	613
Average frequency of a model on the training set			
1280.0	29.9	34.6	30.4
Average number of states per model			
1	1	4.7	3.5
Phonetic accuracy with no language model			
47.0	55.5	81.2	70.6
Word accuracy with a bigram model			
83.9	86.5	82.3	76.1

Table 2.
Comparison of systems based on multigrams and on linguistic units.

Comparison with the reference systems The results obtained with the reference systems and with the multigram system based on 32 acoustic symbols are in Table 2. When no model of language is used, the recognition scores evaluate the reliability of the acoustic modeling within each system, and also, in the multigram case, the reliability of the probabilistic mapping between the sequences of acoustic symbols and of phonemes. The phonetic accuracy obtained with multigram units (70.6 %) is intermediate between the scores obtained with triphones[6] (55.5 %) and with words (81.2 %). The motivation for inferring acoustic multigrams is to find a set of units being altogether the most recurrent and the least variable as possible. Now, assuming that the average variability of recognition units is directly related to their average length[7], words better comply with those requirements than the multigrams derived in our experiment. Indeed, not only the average length of the words, measured by the average number of states per model, but also their average frequency in the training data, are higher than those of the multigram units. The same is true for the multigrams with respect to the triphones, the former tending to be more recurrent and longer units than the latter. Scores turn out to be ranked accordingly.
The use of a bigram language model compensates for the relative lack of reliability of the phoneme and triphone models, which, in this case, outperform the models trained on words, with word accuracies of respectively 83.9 % and 86.5 %. The integration of this language model with the acoustic component based on multigrams, though also quite profitable, does not allow to exceed a 76.1% word accuracy. Since conversely, in our experiments, multigrams used without any language model, give better results than phonemes and triphones, it clearly indicates that the way the multigram level and the linguistic level have been interfaced is not optimal. We identify at least two possible major tracks for improvement. First, the decoding could be performed using (4) rather than (5), so that the output of the recognizer would be determined by summing the likelihood values of multiple acoustic transcriptions, instead of being derived from the single best transcription Z^*_{te}. Second, the joint multigram mapping, which associates a closed list of phonetic sequences to a sequence of acoustic symbols, lacks flexibility, especially when lexical constraints are applied. For instance, it may happen that a sentence cannot be retrieved using only the sequences, which are mapped to an acoustic sequence with a non zero probability. A more flexible approach would be, for instance, to have a production model of phonetic sequences associated to each acoustic sequence, instead of a closed list.

[6] at least when, like in our experiments, no model interpolation, nor tied estimation techniques are used to warrant the reliability of the estimates for the models of triphones.

[7] Of course, it is also related to the number of distinct units among which speech segments are partitionned: triphones are of the same length as phonemes, but they provide a more refined partionning of the data.

5. CONCLUSION

Our work explores the possibility of deriving acoustically consistent units from continuous speech. The units are then matched to variable-length phonetic sequences in a fully unsupervised way, and used for speech recognition. This approach is a step towards a more data-driven acoustic modeling; it aims at eliminating the problem of having to decide a priori on the nature of the recognition units, which, conversely, are derived so as to allow the best fit to the data, as well as the most reliable estimates. In our experiments, the recognition scores obtained with models of acoustic multigrams are intermediate between the performances of systems based on triphones and on words. With the use of a lexicon of words and the integration of a bigram language model, triphones and phonemes outperform the multigram units. Further effort should be put into a representation of lexical entries from *a posteriori* acoustic units.

REFERENCES

[1] C. H. Lee, F. K. Soong and B. H. Juang (1995). *Word recognition using whole word and subword models*, Proceedings of ICASSP 89.

[2] M. Bacchiani, M. Ostendorf, Y. Sagisaka and K. Paliwal (1995). *Unsupervised learning of non-uniform segmental units for acoustic modeling in speech recognition*, Proc. of IEEE Workshop on Automatic Speech Recognition.

[3] F. Bimbot, R. Pieraccini, E. Levin and B. Atal (1995). *Variable-Length Sequence Modeling: Multigrams*, IEEE Signal Processing Letters, vol. 2, n^o 6, June 1995.

[4] S. Deligne, F. Bimbot (1995). *Language Modeling by variable length sequences: theoretical formulation and evaluation of multigrams*, Proceedings of ICASSP 95.

[5] J. Rissanen (1978). *Modeling by shortest data description*, Automatica, vol. 14, pp 465-471.

[6] J. O Ruanaidh, W. J. Fitzgerald (1996). *Numerical Bayesian Methods Applied to Signal Processing*, chapter 2, pp 8-9. Springer-Verlag, New York.

[7] B.S. Atal (1983). *Efficient coding of LPC parameters by temporal decomposition*, Proceedings of ICASSP 83.

[8] S. Deligne, F. Yvon, F. Bimbot (1995). *Variable-length sequence matching for phonetic transcription using joint multigrams*, Proceedings of EUROSPEECH 95.