CONTEXT MODELING IN HYBRID SEGMENT-BASED/NEURAL NETWORK RECOGNITION SYSTEMS

Jan Verhasselt *and *Jean-Pierre Martens*

ELIS, University of Ghent, St.-Pietersnieuwstraat 41, B-9000 Gent, Belgium

ABSTRACT

In this paper, we describe the incorporation of context-dependent models in hybrid Segment-Based/Neural Network speech recognition systems. We present alternative probabilistic frameworks and evaluate them by performing speaker-independent phone recognition experiments on the TIMIT corpus. We compare their recognition performances with that of a context-independent hybrid SB/NN system and with the best published performances on this task.

1. INTRODUCTION

Modeling of subword units in context is a standard technique which increases performance of state-of-the-art HMM recognizers significantly. Relatively simple context-independent hybrid HMM/NN systems were reported to be competitive with more complex context-dependent mixture-of-Gaussians systems [1], but it was shown that such hybrid frame-based systems also benefit from context-modeling [2, 3]. Similarly, it was shown that context-dependent modeling boosts the performance of segment-based (SB) system [4, 5] as well as of hybrid SB/NN systems [6, 7].

Apparently, these context-dependent models can be incorporated in hybrid SB/NN systems in many different ways. More specifically, several alternative probabilistic frameworks can be derived depending on the independency assumptions that are introduced to make the probability estimation feasible. Our goal is to compare the recognition performances of several of these context-dependent hybrid SB/NN frameworks among themselves and relative to the performance of a context-independent hybrid SB/NN system. In order to allow a comparison with other (non-hybrid and/or non-segment-based) recognizers, we report the recognition performances on the well-established phone recognition task on the TIMIT corpus.

2. PROBABILISTIC FRAMEWORKS

Phone recognition can be described as a search for the most likely sequence of phones $\underline{u}$, given the sequence $\underline{\mathbf{x}}$ of acoustic vectors. In segment-based recognizers, variable length segments of the speech signal are mapped to the phones. If the i^{th} phone u_i ends in frame s_i, the segmentation is uniquely specified by the sequence $\underline{s} = s_0, s_1, \dots s_i, \dots, s_{L(\underline{u})}$, where $L(\underline{u})$ is the number of phones in $\underline{u}$. The recognition process can then be formally described as a search for:

$$\hat{\underline{u}} = \underset{\underline{u}}{\operatorname{argmax}} \sum_{\underline{s}} Pr(\underline{u}, \underline{s} | \underline{\mathbf{x}}) \tag{1}$$

The most straightforward probabilistic framework is obtained by writing the joint probability $Pr(\underline{u}, \underline{s}|\underline{\mathbf{x}})$ as the product of a *segmentation probability* $Pr(\underline{s}|\underline{\mathbf{x}})$ and a *unit classification probability* $Pr(\underline{u}|\underline{s}, \underline{\mathbf{x}})$ [8]. In [9], we have described the advantages of this decomposition. As we can never hope to estimate the probabilities of these complete sequences in one piece, we must write them as products of the probabilities of their components. Furthermore, we have to introduce approximations to make the estimation feasible. E.g. if context-independent segment and unit classification models are used, $Pr(\underline{u}, \underline{s}|\underline{\mathbf{x}})$ is approximated as:

$$\prod_{i=1}^{L(\underline{u})} Pr(s_i | s_{i-1}, \mathbf{Y}_i; \lambda_s) Pr(u_i | s_i, s_{i-1}, \mathbf{Z}_i; \lambda_u) \tag{2}$$

where λ_s and λ_u denote the parameters of the segment model and the unit classification model respectively. This notation makes explicit that it are probability *estimates*, made by a certain model, rather than true probabilities. In our hybrid SB/NN system, both probabilities are estimated by Multi-Layer Perceptrons (MLP's), and the $\lambda' s$ simply denote the weights of the respective MLP's. $\mathbf{Y}_i(\underline{\mathbf{x}})$ and $\mathbf{Z}_i(\underline{\mathbf{x}})$ represent segmental feature vectors. A segmental feature vector is the result of a transformation $f(\underline{\mathbf{x}}, s_{i-1}, s_i)$ of the acoustic vector sequence $\underline{\mathbf{x}}$ with the purpose of describing the variable length segment $]s_{i-1}, s_i]$ in its acoustic context. The latter means that the transformation can take acoustic vectors $\mathbf{x}$ from outside the segment $]s_{i-1}, s_i]$ into account. Note that this is allowed since the whole sequence $\underline{\mathbf{x}}$ occurs in the conditioning parts of equation 1. In section 4, we will present some details of this neural modeling.

In case a left-phonetic context dependent unit classification model is used, $Pr(\underline{u}, \underline{s}|\underline{\mathbf{x}})$ is approximated as:

$$\prod_{i=1}^{L(\underline{u})} Pr(s_i | s_{i-1}, \mathbf{Y}_i; \lambda_s) Pr(u_i | u_{i-1}, s_i, s_{i-1}, \mathbf{Z}_i; \lambda_c) \tag{3}$$

where λ_c denotes the parameters of the context-dependent model.

The probabilistic frameworks specified by equation 2 and equation 3 both have the same drawback: they do not allow the explicit use of a language model λ_l, or, in other words: they implicitly use the "language model" that the unit classification MLP's have learned from the training corpus: in case of the CI models this is the unigram language

[*] Aspirant F.W.O. – Belgacom

model on the phone level and in case of the left-context dependent models this is the bigram language model on the phone level.

In order to allow the explicit use of a language model (which is particularly useful for word recognition, but, as we will soon see, also for phone recognition), equation 1 is usually rewritten as:

$$\underline{\hat{u}} = \underset{\underline{u}}{\operatorname{argmax}} \sum_{\underline{s}} Pr(\underline{u}) \frac{P(\underline{s}, \underline{\mathrm{x}}|\underline{u})}{P(\underline{\mathrm{x}})} \tag{4}$$

This approach is also used in non-hybrid segment-based systems, in which case $Pr(\underline{u})$ is modeled by an explicit language model $Pr(\underline{u}|\lambda_l)$, whereas $P(\underline{s}, \underline{\mathrm{x}}|\underline{u})$ is modeled by the "acoustical" model $P(\underline{s}, \underline{\mathrm{x}}|\underline{u}; \lambda_a)$. The denominator $P(\underline{\mathrm{x}})$ is usually ignored during decoding since it is independent of $\underline{u}$ and $\underline{s}$, resulting in:

$$\underline{\hat{u}} = \underset{\underline{u}}{\operatorname{argmax}}\, Pr(\underline{u}|\lambda_l) \sum_{\underline{s}} P(\underline{s}, \underline{\mathrm{x}}|\underline{u}; \lambda_a) \tag{5}$$

In order to derive a framework that is suited for models that estimate posterior probabilities (such as MLP's) equation 5 can be rewritten using Bayes' law as:

$$\underline{\hat{u}} = \underset{\underline{u}}{\operatorname{argmax}}\, Pr(\underline{u}|\lambda_l) \sum_{\underline{s}} \frac{Pr(\underline{s}|\underline{\mathrm{x}}; \lambda_a) Pr(\underline{u}|\underline{s}, \underline{\mathrm{x}}; \lambda_a)}{Pr(\underline{u}|\lambda_a)} \tag{6}$$

where $P(\underline{\mathrm{x}}|\lambda_a)$ has been ignored for the same reasons as $P(\underline{\mathrm{x}})$ above.

Again, modeling assumptions are necessary in order to make the estimation of these probabilities feasible. E.g. in case context-independent acoustical models are used, the recognition process is a search for:

$$\underline{\hat{u}} = \underset{\underline{u}}{\operatorname{argmax}} \sum_{\underline{s}} \prod_{i=1}^{L(\underline{u})} Pr(u_i|u_{i-1}, .., u_1, \lambda_l) \cdot Pr(s_i|s_{i-1}, \mathbf{Y}_i; \lambda_s) \frac{Pr(u_i|s_i, s_{i-1}, \mathbf{Z}_i; \lambda_u)}{Pr(u_i|\lambda_u)} \tag{7}$$

This is the probabilistic framework that we have used in our Context-Independent Discriminative Stochastic Segment Model (DSSM) hybrid SB/NN system [9, 10] with a language model that is an interpolation between the unigram and bigram language model on the phone level:

$$Pr(\underline{u}|\lambda_l) = \prod_{i=1}^{L(\underline{u})} \alpha Pr(u_i|u_{i-1}; \lambda_l) + (1-\alpha) Pr(u_i|\lambda_l) \tag{8}$$

In the remainder of this paper, we will refer to this language model as "α-bigram". If $\alpha = 1$, it reduces to a bigram language model, and if $\alpha = 0$, equation 7 reduces to equation 2, at least if the unit classification model has well learned the unigram probability from the training corpus, i.e. $Pr(u_i|\lambda_u) = Pr(u_i|\lambda_l)$. This probabilistic framework is familiar with the one used in the Segmental Neural Net (SNN) system [7]. However, in the latter the segmentation probability is not incorporated explicitly. Instead, the SNN is trained with a so called *N-Best training* procedure. In [9], we show that this N-best training procedure has a similar functionality as the segmentation probability.

In [6], a different probabilistic framework for incorporating context in hybrid SB/NN systems is presented. In that paper, the framework is derived for word recognition. However, phone recognition can be considered as a special case of word recognition, where the "words" w to recognize are nothing else than the phones u. Therefore, using the substitution $Pr(\underline{w}) \rightarrow Pr(\underline{u}|\lambda_u)$, equation (5) in [6] can be written as:

$$\underline{\hat{u}} = \underset{\underline{u}}{\operatorname{argmax}}\, Pr(\underline{u}|\lambda_l) \sum_{\underline{S}} \prod_{i=1}^{L(\underline{u})} \frac{Pr(\gamma_i, S_i|\mathbf{X}_i)}{Pr(\gamma_i)} \tag{9}$$

where γ_i represents a triphone consisting of (u_{i-1}, u_i, u_{i+1}) and S_i represents a phonetic segment (in [6], the segmentation is represented as a sequence of phonetic segments).

In the following, we will translate equation 9 to our notation with the purpose of comparing it with the frameworks derived above. As we represent the segmentation as a sequence of phonetic boundaries, S_i is translated as (s_{i-1}, s_i). Furthermore, we generalize equation 9 in order to represent an arbitrary context model by writing the triphone as (u_i, c_i), with context $c_i = (u_{i-1}, u_{i+1})$.

$$\underline{\hat{u}} = \underset{\underline{u}}{\operatorname{argmax}} \sum_{\underline{s}} Pr(\underline{u}|\lambda_l) \prod_{i=1}^{L(\underline{u})} \frac{Pr(u_i, c_i, s_i, s_{i-1}|\mathbf{X}_i)}{Pr(u_i, c_i)} \tag{10}$$

If context-independent models are used (i.e. if $c_i = \phi$), equation 10 reduces to:

$$\underline{\hat{u}} = \underset{\underline{u}}{\operatorname{argmax}}\, Pr(\underline{u}|\lambda_l) \sum_{\underline{s}} \prod_{i=1}^{L(\underline{u})} Pr(s_{i-1}|\mathbf{X}_i; \lambda_b) \cdot Pr(s_i|s_{i-1}, \mathbf{X}_i; \lambda_s) \frac{Pr(u_i|s_i, s_{i-1}, \mathbf{X}_i; \lambda_u)}{Pr(u_i|\lambda_u)} \tag{11}$$

where λ_b denotes the parameters of the boundary model (in our system the weights of an MLP estimating $Pr(s_{i-1}|\mathbf{X}_i)$.

Note that the only difference between equation 11 and equation 7 is the presence of the boundary probability estimate $Pr(s_{i-1}|\mathbf{X}_i; \lambda_b)$.

Similarly, if diphone models are used (i.e. $c_i = u_{i-1}$), equation 10 reduces to:

$$\underline{\hat{u}} = \underset{\underline{u}}{\operatorname{argmax}}\, Pr(\underline{u}|\lambda_l) \sum_{\underline{s}} \prod_{i=1}^{L(\underline{u})} Pr(s_{i-1}|\mathbf{X}_i; \lambda_b) \cdot Pr(s_i|s_{i-1}, \mathbf{X}_i; \lambda_s) \frac{Pr(u_{i-1}|s_i, s_{i-1}, \mathbf{X}_i; \lambda_u)}{Pr(u_{i-1}|\lambda_u)} \cdot \frac{Pr(u_i|u_{i-1}, s_i, s_{i-1}, \mathbf{X}_i; \lambda_c)}{Pr(u_i|u_{i-1}; \lambda_c)} \tag{12}$$

As we have not yet trained an MLP in order to estimate the probability $Pr(u_{i-1}|s_i, s_{i-1}, \mathbf{X}_i; \lambda_u)$, we were not able to implement equation 12 completely. However, assuming that the phonetic identity of the previous phone u_{i-1} can not be learned from the acoustic observations $\mathbf{X}_i$ and the duration $d_i = s_i - s_{i-1}$ of the present segment i (i.e. assuming $Pr(u_{i-1}|s_i, s_{i-1}, \mathbf{X}_i; \lambda_u) = Pr(u_{i-1}|\lambda_u)$, and using a bigram language model $Pr(\underline{u}|\lambda_u) = Pr(u_i|u_{i-1}; \lambda_c)$, equation 12 reduces to equation 3 except for the additional boundary probability $Pr(s_{i-1}|\mathbf{X}_i; \lambda_b)$.

3. THE DSSM RECOGNITION SYSTEM

The various probabilistic frameworks presented in the previous section (equations 2,3,7, 11, and 12) are evaluated in our DSSM system [10, 9]. This system incorporates an auditory model front-end, a presegmentation algorithm, Multi-Layer Perceptrons (MLP's) for the computation of boundary, segment and (context-independent or -dependent) unit probabilities, a lexical network, and a search and decision module. The search can be either a full forward search (for phone recognition, this implies effectively summing over $\underline{s}$ in the previous equations) or alternatively a Viterbi search in which the sum is replaced by a maximum operator.

4. NEURAL MODELING

All posterior probabilities are estimated by fully connected MLP's 2-layer MLP's with sigmoidal activation functions. They are trained by error backpropagation, using a least mean squares cost function. The "correct" segmentation and phone sequence was obtained from the manual annotation that is supplied with the TIMIT corpus.

The context-independent posterior phone probability estimate $Pr(u_i|s_i, s_{i-1}, \mathbf{Z}_i; \lambda_u)$ is the weighted average of the corresponding outputs of three MLP's with a total number of 460900 weights:

$$Pr(u_i|s_i, s_{i-1}, \mathbf{Z}_i; \lambda_u) = \sum_{j=1}^{3} g_j Pr(u_i|s_i, s_{i-1}, \mathbf{Z}_i^j; \lambda_u^j) \quad (13)$$

The optimal interpolation coefficients g_j are obtained by solving a linear matrix equation. Each of these so called "unit" MLP's has a different input vector, consisting of a fixed length segmental feature vector $\mathbf{Z}_i^j$ and the segment duration d_i. The segmental features are averages of acoustic vectors in sub-segments, selected acoustic vectors, correlation coefficients and time derivatives of the acoustic parameters, maximal energy in the segment and voicing evidence. More details about the segmental features can be found in [9]. Each MLP has an output node for each of the phones. These MLP's are trained on all the phonetic segments in the training corpus. The teaching outputs were 1 for the correct phone and 0 for all the other phones.

The probability estimate $Pr(u_i|u_{i-1}, s_i, s_{i-1}, \mathbf{Z}_i; \lambda_c)$ is estimated as a weighted average of the outputs of three MLP's with a total number of 508885 weights. Each MLP has an output node for each of the phones and an input vector consisting of $\mathbf{Z}_i^j$ and d_i, just like the corresponding unit MLP. However, these left-context dependent unit MLP's have additionally a number of inputs (one for each phone) specifying the phonetic identity of the left-context u_{i-1}: the context input corresponding to the left phone is 1 and all other context inputs are 0. This one-of-n encoding results in a higher classification accuracy than binary encoding.

The segment probability estimate $Pr(s_i|s_{i-1}, \mathbf{Y}_i; \lambda_s)$ is a weighted average of the outputs of three so-called "segment" MLP's with a total number of 13278 weights. The input vectors consist of the segmental feature vectors $\mathbf{Y}_i^j$ and of d_i. Each network has one output and is trained on all segments that can be hypothesized during the recognition process and that start on a true phone boundary. The latter is required because the previous phone boundary s_{i-1} occurs in the conditioning part. The teaching output is 1 for phonetic segments and 0 for all others.

The boundary probability estimate $Pr(s_{i-1}|\mathbf{X}_i; \lambda_b)$ is the output of a "boundary" MLP with 2951 weights that is trained on all the potential phone boundary locations in the training corpus. The teaching output is 1 for locations corresponding to a phone boundary, and 0 for all others.

5. EXPERIMENTS AND RESULTS

The experiments deal with phone recognition on the American English corpus TIMIT. Every 10ms, an acoustic parameter vector x consisting of an auditory spectrum, a voicing evidence and a total energy is generated. The experimental results reported here are for the 39 phones set defined in [11]. All unit MLP's were trained for these 39 phones, and for the glottal stop. The language model was also specified in terms of these 40 phones. For the evaluation, the glottal stop was removed from both the recognized and the manual phone sequence. Training was performed on the NIST designated training corpus (3696 sentences), and testing on the core test corpus (192 sentences). Some experiments were repeated on the full test corpus (1344 sentences). System development was performed on a subset of the full test corpus (320 sentences) that contains no sentences from the core test corpus. The "sa" sentences were excluded from the training, development and testing sets.

5.1. Comparison of probabilistic frameworks

Table 1 shows the total phone recognition error rates (TE) (deletion + insertion + substitution errors) of the probabilistic frameworks derived above (using a Viterbi recognition search): It can be observed that the diphone segment model (using equation 3) obtains about the same recognition performance as the α-bigram context-independent model, in spite of the fact that the diphone models obtain a higher classification performance (79.3% correct) than the context-independent models (77.5% correct) on the test set. Possibly, the diphone models will be better capable of exploiting the increased modeling capacity that becomes available if the number of weights is further increased (for the experiments reported here, we used about the same number of weights for both models). We intend to verify this hypothesis in the near future.

Eqn.	Description	core	full
2	implicit unigram, CI SM	30.8	/
3	implicit bigram, Diphone SM	30.5	29.9
7	α-bigram, CI SM	30.4	29.9
11	α-bigram, CI SM	30.8	/
12	bigram, Diphone SM	31.2	/
3+7	(α)-bigram, CI+Diphone SM	29.8	29.4

Table 1: Total phone recognition error rates (TE) [%] on the TIMIT core and full test set. CI = context-independent, SM = segment model, Eqn. = equation number.

Furthermore, using the same neural models, the probabilistic frameworks derived from equation 10 yield a lower recognition performance than those derived from equation 7. This indicates that the boundary probability hurts the recognition performance (remember that we were only able to implement a reduced version of equation 12 in time).

The last line in the table shows the error rate for the combination of the α-bigram CI segment model (equation 7) with the implicit bigram CD segment model (equation 3). The two systems are combined by linearly interpolating their respective probability estimates. The improvement of the recognition performance shows that the two systems are to some extent complementary.

5.2. Comparison with published results

For comparing the best DSSM systems with the best published results on the TIMIT core test corpus, we have used a full forward rather than a Viterbi search, as this yields a slightly higher recognition performance. The results are presented in Table 2. No distinction is made between α-bigram and true bigram, as it is not clear if interpolation is used in the published systems. A direct comparison is somewhat difficult due to differences regarding the complexity of the acoustic and language models. Nevertheless, we believe that our results are competitive with those obtained by others.

Reference	Description	TE [%]
[12]	Bigram, CI STM	36.0
[12]	Trigram, Triphone STM	30.5
[13]	Bigram, Triphone CDHMM	30.9
[14]	Bigram, CI 2nd order HMM	31.2
[1]	Bigram, Recurrent Network	26.6
[15]	Bigram, CI SM	35.9
[15]	Bigram, Diphone SM	30.5
[5]	Trigram, Diphone frame+SM	26.6
Eqn.7	Bigram, CI SM	30.3
Eqn.3	Bigram, Diphone SM	30.5
Eqn.3+7	Bigram, CI + Diphone SM	29.7

Table 2: Published total phone recognition error rates (TE) on the TIMIT core test set. CI = context-independent, SM = segment model, Eqn. = equation number.

6. CONCLUSION

We have evaluated different probabilistic frameworks that allow the incorporation of context-dependent modeling in hybrid Segment-Based/Neural Network speech recognition systems. Even for the best frameworks, a diphone-based system does not obtain a higher recognition performance than a context-independent (CI) system that uses a bigram language model, if the number of model parameters is kept fixed. Nevertheless, the combination of the diphone- and CI-based system does yield a higher recognition performance. We believe that our results are competitive with those obtained by others.

7. REFERENCES

[1] A.J. Robinson, "An Application of Recurrent Nets to Phone Probability Estimation," *IEEE Trans. on Neural Networks*, Vol. 5, no. 2, pp. 298-305, 1994.

[2] H. Franco, M. Cohen, N. Morgan, D. Rumelhart, and V. Abrash, "Context-Dependent Connectionist Probability Estimation in a Hybrid Hidden Markov Model - Neural Net Speech Recognition System," *Computer Speech and Language*, Vol. 8, No. 3, pp. 211-222, 1994.

[3] J. Fritsch, M. Finke, A. Waibel, "Context-Dependent Hybrid HME/HMM Speech Recognition using Polyphone Clustering Decision trees," *Procs ICASSP*, Vol. 3, pp. 1759-1762, 1997.

[4] O. Kimball, "Segment Modeling Alternatives for Continuous Speech Recognition," *PhD. Thesis*, Boston University College of Engineering, 1995.

[5] J. Chang, and J. Glass, "Segmentation and Modeling in Segment-Based Recognition," *Procs of EUROSPEECH*, Vol. 3, pp. 1199-1202, 1997.

[6] H. Leung, I. Hetherington, and V. Zue, "Speech Recognition using Stochastic Segment Neural Networks," *Procs of ICASSP*, Vol. 1, pp. 613-616, 1992.

[7] G. Zavaliagkos, Y. Zhao, R. Schwartz, and J. Makhoul, "A Hybrid Segmental Neural Net/Hidden Markov Model System for Continuous Speech Recognition," *IEEE Trans. on Speech and Audio Proc.*, Vol. 2, No. 1, Part II, pp. 151-160, 1994.

[8] M. Ostendorf, V. Digalakis, and O. Kimball, "From HMM's to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 5, pp. 360-378, 1996.

[9] J. Verhasselt, I. Illina, J.-P. Martens, Y. Gong, and J.-P. Haton, "Assessing the Importance of the Segmentation Probability in Segment-Based Speech Recognition," *To appear in Speech Communication.*

[10] J. Verhasselt, J.-P. Martens, and B. Baeyens, "Speech Recognition Using a Discriminative Context-Independent, Segment-Based Speech Recognizer," *Procs of IEEE ProRISC*, pp. 367-372, 1996.

[11] K.F. Lee, and H.W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models," *IEEE Trans. on ASSP*, 37 (11), 1989.

[12] W. Goldenthal, "Statistical Trajectory Models for Phonetic Recognition", PhD. Thesis, Laboratory for Computer Science MIT, 1994.

[13] L. Lamel and J.L. Gauvain, "High performance Speaker-Independent Phone Recognition using CDHMM," *Procs of EUROSPEECH*, Vol. 1, pp. 121-124, 1993.

[14] J.F. Mari, D. Fohr, and J.C. Junqua, "A Second-Order HMM for High Performance Word and Phoneme-Based Continuous Speech Recognition," *Proc. of ICASSP*, Vol. 1, pp. 435-438, 1996.

[15] J. Glass, J. Chang, M. McCandless, "A Probabilistic Framework for Feature-Based Speech Recognition," *Proc. of ICSLP*, Vol. 4, pp. 2277-2280, 1996.