ACOUSTIC-PHONETIC DECODING USING A TRANSITION CONTROLLED NEURAL NET

Jan P. Verhasselt[*] *and Jean-Pierre Martens*[†]

ELIS, University of Ghent
St.-Pietersnieuwstraat 41, B-9000 Gent (Belgium)
Jan.Verhasselt@elis.rug.ac.be

ABSTRACT

In this paper, an artificial neural network (ANN) architecture for modeling the transitions between consecutive phones is presented. These 'phone transition' models are particularly suited for taking into account the coarticulation phenomena in continuous speech. In order to obtain robust and generalizing probability estimates, the evidences of variable frame rate-based transition models and those of context-independent segment-based phone models are combined by means of an additional ANN, called the Transition Controlled Neural Network (TCNN). The concept of the transition approach was already introduced in [1], but in this paper a new and more sophisticated implementation is proposed and evaluated on a phone recognition task. The new TCNN-approach significantly outperforms the old one.

1. INTRODUCTION

Although the aim of acoustic-phonetic decoding is to transform the acoustic continuum of speech into a sequence of discrete linguistic units, the speech signal is not a sequence of independent discrete events. This implies that the acoustical realization of a phone is partly determined by the identities of the preceding and the following phone (and, in general to a smaller extent, by the further context). Moreover, some phones, such as voiced stops [2] and nasals [3], are preferably described by formant transitions towards the phone and away from the phone.

A common way of dealing with these coarticulation phenomena, is by using context-dependent phone models. Many successful implementations have been reported: frame-based [4, 5] as well as segment-based [6, 7], and connectionist [5, 6] as well as non-connectionist [4, 7]. A potential difficulty with context-dependent (phone) models is the reduced amount of data available to train them. Therefore, many systems perform an interpolation between models trained at different levels of context specificity. In order to simplify this interpolation, the context-dependent and independent models usually are of the same type (all segment-based or all frame-based, and all connectionist or all non-connectionist) and use the same observation vectors.

Since the different model-types have specific advantages and drawbacks, it is appealing to combine them in a way that the various advantages are retained as much as possible, whereas most of the drawbacks are eliminated. In this way, specialized models, with dedicated acoustic feature vectors, could be used at the various levels of context-specificity. We already proposed [1] a framework in which segment-based context-independent phone models are combined with variable frame rate-based transition models.

It has been argued in many papers [1, 6, 7, 8, 9, 10] that a segment-based approach could offer definite advantages over the frame-based approach. In particular, we mention the flexibility of introducing segment-related prior knowledge about speech, and the capability of capturing the spectral/temporal relationships over the whole phone.

However, the context-independent phone models are not able to deal satisfactorily with coarticulation phenomena. A possible solution would be to introduce context-dependent segment-based phone models. However, the segment-based approach exhibits some practical difficulties:

1. The classification models assume a given segmentation. Since it is impossible to determine the correct segmentation before classification, a large number of reasonable segmentations have to be considered during the dynamic search. While this remains feasible with a limited number of context-independent phone models, it becomes computationally expensive when a large number of context-dependent models have to be investigated in those hypothesized segments.

2. The observation vectors observed in a segment have to be warped to a fixed length vector that can be analyzed by the classification models. This warping may delete or obscure relevant information.

In order to avoid these difficulties, we decided to adopt another approach. We designed phone transition models that examine the spectral transitions in a confined region surrounding a potential boundary between two phones. These models are supplied with dedicated observation features that capture the typical dynamics in the spectral balance, the energy and the voicing evidence in those regions. In [1], we used non-connectionist transition models: difference measures between the actual acoustical observations and a 'template' that is typical for the hypothesized phone-pair. In this paper, we introduce connectionist transition models.

The scores of the segment-based context-independent phone models and the transition model scores are provided as inputs to an additional ANN, called the Transition Controlled Neural Network (TCNN). This TCNN transforms

[*] Aspirant N.F.W.O. – Belgacom

[†] Research Associate of the National Fund for Scientific Research

0-7803-2431-5/95 $4.00 © 1995 IEEE

these scores into probability estimates that are suited for a variable frame rate dynamic search [11]. Thus the interpolation between the transition models and the more general and better trained phone models is performed using a discriminant procedure, instead of a likelihood maximizing procedure such as deleted interpolation [12].

There exist other approaches in the literature which explicitly model the phonetic transitions [7, 8]. These approaches use a joint-Gaussian distribution to model the error between the observations and a 'track' or 'template' that is typical for the transition. There are three major differences between these approaches and ours:

1. We use connectionist transition models, in addition to 'template' or 'track' models.
2. The transition models in [8] examine regions which are related to the hypothesized phonetic segments. The transition models in [7] examine a fixed number of frames around the boundary. In our transition models too, part of the observation features describe a fixed number of frames around the boundary, and are thus unwarped. Other features describe variable length segments, emerging from an initial segmentation (see next section).
3. The dynamic search in our approach is the efficient variable frame rate search, instead of the much more computationally expensive stochastic segment search.

2. CONTEXT-INDEPENDENT PHONETIC RECOGNITION

The context-independent phone models were provided by the phonetic module of the segment-based phone recognition system described in [10]. This system incorporates an auditory model front-end, an initial segmentation stage and a phonetic classification and segmentation module. The auditory model generates a sequence $\overline{o}$ of N_f observation vectors each characterizing a 10 ms speech frame. The initial segmentation module retains a set $\overline{b}$ of $N_b < N_f + 1$ candidate phonetic segment boundaries. A phonetic segment boundary is defined as a boundary between the acoustic realizations of subsequent phones. The segments of speech enclosed by two consecutive candidate phonetic boundaries are called 'initial segments'. Candidate phonetic segments are built by concatenating up to four consecutive initial segments, and a Multi Layer Perceptron (MLP) is trained to estimate the posterior probability that a candidate phonetic segment is really phonetic. Other MLP's are trained to estimate the posterior probabilities of particular phones u_j ($j = 1..K$) being realized in those phonetic segments. The inputs of these MLP's are derived from the observation vectors in the segment and its close surroundings. They include segmental features such as the duration of the segment.

The Dynamic Programming (DP) search examines several candidate phonetic segmentations $\overline{s}$ and phone sequences $\overline{u}$ of the same length as $\overline{s}$, and calls the MLP's to determine the probabilities of these phonetic decodings ($\overline{s}$,$\overline{u}$), given the acoustic evidence.

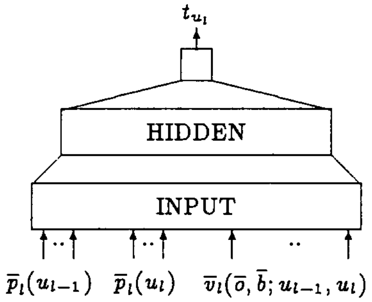
t_{u_l}
HIDDEN
INPUT
$\overline{p}_l(u_{l-1})$
$\overline{p}_l(u_l)$
$\overline{v}_l(\overline{o},\overline{b};u_{l-1},u_l)$

Figure 1: TCNN-structure.

3. THE TRANSITION APPROACH

In our transition approach, the acoustic continuum of speech is described as a sequence of transitions between consecutive phones. If those transitions are assumed to occur on the boundaries in $\overline{b}$, the phonetic decoding task can be described as one of finding a sequence $\overline{t}$ of N_b transitions between phone-pairs, and $\overline{s}$ and $\overline{u}$ can be derived from $\overline{t}$. The transition models should estimate the probability of observing a transition t_{u_l} to a particular phone u_l in the vicinity of the boundary b_l being analyzed, given that u_{l-1} was estimated at the previous boundary b_{l-1}, and given the acoustic observations. Obviously, as one must be able to deal with inserted (phone internal) candidate boundaries, transitions between two identical phones (in fact parts of a phone) have to be investigated as well (i.e. $u_{l-1} = u_l$). This is a direct consequence of using a variable frame rate-based search, instead of a segment-based search.

In [1], we introduced consecutive approximations in the probabilistic framework, in order to reduce the amount of free parameters that have to be estimated from the training corpus. This reduction is necessary in order to assure the trainability of the nets. In this paper, we will only discuss the TCNN framework and structure; the interested reader can find more details in [1].

3.1. The TCNN approach

We factorized and approximated the posterior probability $P(\overline{t}|\overline{o},\overline{b})$ of $\overline{t}$, given the acoustical observations and the initial segmentation as follows [1]:

$$\prod_{l=1}^{N_b} P(t_{u_l}|u_{l-1},\overline{v}_l(\overline{o},\overline{b};u_{l-1},u_l),\overline{p}_l(u_{l-1}),\overline{p}_l(u_l)) \qquad (1)$$

In this expression, $\overline{p}_l(u_l)$ represents context-independent posterior probabilities of the phone u_l in some candidate phonetic segments in the vicinity of b_l, and $\overline{v}_l(\overline{o},\overline{b};u_{l-1},u_l)$ represents the scores of the various transition models, indicating the correspondence (or difference) between the actual acoustical observations and those which are typical for the investigated transition. These transition models will be discussed in the following section. The probabilities in equation (1) are estimated by the MLP depicted in figure 1. This MLP, called the 'Transition Controlled Neural Network', has only one output node, and estimates the posterior probability of the hypothesized transition. By using the TCNN in a hypothesis testing scheme, we only need to provide phone probabilities for the two phones of the dyad being investigated. In this way, the number of inputs is restricted, but the TCNN must be called at each potential boundary as many times as there are transitions to examine. By examining only those transitions corresponding with the $u_{l-1}'s$ of the most promising paths to b_l (b_l not included), and with the $u_l's$ receiving sufficient evidence from the context-independent module, the computational load is reduced, whereas the recognition performance is not affected.

Note that the proposed TCNN contains no nodes which are phone-pair specific, nor is the identity of the hypothesized dyad specified to the TCNN. Consequently, the TCNN itself is not able to learn the prior transition probabilities from the training set. In fact, all features constructed by the TCNN are shared by all dyads and everything that is really phone-pair specific has to be captured in the TCNN inputs. This explains why the phone probabilities $\overline{p}_l(u_{l-1})$ and $\overline{p}_l(u_l)$, as well as the transition models scores $\overline{v}_l(\overline{o},\overline{b};u_{l-1},u_l)$ are given as TCNN inputs: they all give an indication of the correctness of the hypothesis.

3.2. Transition models

3.2.1. Difference Measures

In [1], we used extremely simple transition models : difference measures quantifying the differences between the actual acoustical observations $(\overline{o},\overline{b})$ observed in the vicinity of b_l, and those which are typical for (u_{l-1},u_l). In order to limit the number of free parameters, the statistical distribution of a feature $x_m(\overline{o},\overline{b})$ for a particular phone-pair (u_{l-1},u_l) is characterized by the maximum likelihood estimates of its mean $\mu_m(u_{l-1},u_l)$ and its standard deviation $\sigma_m(u_{l-1},u_l)$. The corresponding difference measure which serves as an input to the TCNN is then given by:

$$v_{lm}(\overline{o},\overline{b};u_{l-1},u_l)=\frac{|x_m(\overline{o},\overline{b})-\mu_m(u_{l-1},u_l)|}{\sigma_m(u_{l-1},u_l)} \qquad (2)$$

No prior assumptions (such as a particular parametric form, e.g. multivariate Gaussian) about the statistical distributions of the features are made, except that they are expected to be more or less symmetrical around their means. It is left to the Error Backpropagation (EBP) training of the TCNN to decide on how to combine the context-independent probabilities and the context-dependent difference measures.

Although this TCNN with Difference Measures as inputs (DM-TCNN) yielded a significant improvement over the context-independent system [1], the rather primitive way of characterizing the feature-distributions, and the fact that the TCNN is not able to learn the prior transition probabilities do put an upper limit on the attainable performance. A better description of the template could be made by estimating the full covariance matrix. But as we have unsufficient data to estimate this large number of free parameters, we would have to cluster several transition models or to make some independency assumptions. Instead, we implemented a totally different kind of transition models: MLP's with different degrees of context-specificity were trained to estimate the posterior probabilities $P(t_{u_l}|c_{l-1},\overline{o},\overline{b})$ of the transitions, with c_{l-1} specifying the identity of u_{l-1} to some degree. These probabilities are then supplied to the TCNN in addition to the context-independent phone probabilities and the former difference measures. In the following section, we will describe the implementation of these connectionist transition models.

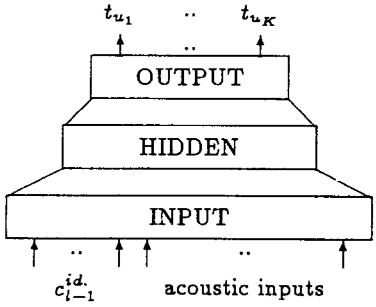

Figure 2: Structure of the connectionist transition models. The $c^{id.}_{l-1}$ inputs specify the identity of u_{l-1}.

3.2.2. Connectionist Transition Models

Currently, we implemented two connectionist transition models: one that is conditioned on the preceding phone (i.e. $c_{l-1}=u_{l-1}$, 'biphone type'), and one that is conditioned on the Broad Phonetic Class (BPC : front vowel, center vowel, back vowel, sonorant, unvoiced fricative, voiced fricative, stop and silence) of that phone (i.e. $c_{l-1}=$ BPC of u_{l-1}, 'generalized biphone type'). This way, a significant sharing of training examples is accomplished. Since we have insufficient data to train a MLP for every phone-pair, the biphone type transition model is implemented as a single MLP having an output for every right phone and having the identity of the left-context explicitly supplied at its inputs by a one-of-n encoding (figure 2). This way, the weights of the connections between the inputs and the hidden layer are shared among all the phone-pairs, and the weights of the connections from the hidden layer to a particular output-node are shared among all phone-pairs with the same right-phone. This network can learn the prior transition probabilities from the training set, since it is able to capture the correlations between the one-of-n input patterns (left-phone) and the desired output patterns (right-phone). Similarly, the generalized biphone type transition model is implemented as another MLP, with a one-of-n encoding of the left-BPC. Obviously, both MLP's are supplied with the usual inputs describing the acoustical observations in the neighborhood of the investigated boundary. Such network structures are less sensitive to the problem of sparse data, since they can interpolate between context-dependent and independent models [6].

4. CORPUS AND TRAINING

The training corpus consisted of 780 phonetically balanced Dutch sentences (25 minutes of continuous speech) originating from 60 speakers. The training of the context-independent module is described in [10]. The training utterances were automatically aligned to their phonetic transcriptions, and the means and the variances characteristic for the dyads were computed from these alignments. For every candidate phonetic boundary b_l, input vectors for the connectionist transition models were stored in a training database, accompanied by two labels indicating the identity of the phone to the left and to the right of the boundary. Both networks were trained on these databases using the EBP algorithm. In a second alignment, the inputs of the TCNN (including the outputs of the two connectionist transition models) were stored in a training database. The training is completed by training the TCNN on this database, again using the EBP algorithm.

5. EXPERIMENTAL RESULTS

A separate test set of 130 sentences from 10 new speakers was hand-segmented and labeled for evaluation. Both the training and the test set were recorded in a noise free room. The TCNN-based system was evaluated on a speaker-independent phone recognition task, and compared with the stochastic segment MLP/DP hybrid which also provided the context-independent phone probabilities. The results obtained on the test set are displayed in table 1. Phone recognition results are shown for the context-independent baseline system, the DM-TCNN, and the new TCNN. This new TCNN implicitly uses a bigram grammar, through the posterior probability estimate of the biphone type connectionist transition model. The TCNN-based system realizes an improvement over the baseline system which is 95% significant.

If only the output of the biphone type connectionist transition model is used in the variable frame rate search (instead of the TCNN-output), the total error is 47.2%. Similarly, if only the generalized biphone type connectionist transition model is used, the total error is 48.7%. If the TCNN is only supplied with the context-independent phone probabilities, the total error is 44.3%, which is worse than the segment-based context-independent system. We conjecture that this performance loss is due to the less optimal variable frame rate search. When only the transition scores are supplied to the TCNN, the total error is 40.9%. The best performance is obtained when the TCNN combines the scores off all models.

	Baseline system		DM-TCNN	TCNN
	Unigram	Bigram	Unigram	Bigram
D	11.4%	11.4%	10.0%	10.7%
I	4.8%	4.1%	6.1%	4.3%
S	25.2%	24.4%	25.0%	22.3%
T	41.4%	39.9%	41.1%	37.3%

Table 1: Phone Recognition Results : D = deletions, I = insertions, S = substitutions, T = total error.

6. CONCLUSION

The TCNN-based system proposed in this paper significantly outperforms a context-independent system. This indicates that transition models can effectively capture coarticulation phenomena in speech and that the TCNN can interpolate the context-dependent transition evidences with the context-independent segment-based phone evidences in an appropriate way.

7. ACKNOWLEDGEMENT

This work was partly performed under the N.F.W.O.-IBM Grant 1993-1994.

8. REFERENCES

[1] J. Verhasselt, J.-P. Martens, "Phone Recognition using a Transition-Controlled, Segment-based DP/MLP Hybrid," in *Procs ICSLP*, vol. 3, pp. 1495-1498, 1994.

[2] J. Olive, A. Greenwood, and J. Coleman, "Acoustics of American English Speech: a Dynamic Approach," New York Berlin Heidelberg: Springer-Verlag, 1993.

[3] G. Fant, "Acoustic Description and Classification of Phonetic Units," in *Speech Sounds and Features*, MIT Press, 1973.

[4] R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner, and J. Makhoul, "Context-dependent modeling for acoustic-phonetic recognition of continuous speech," in *Procs ICASSP*, pp. 1205-1208, 1985.

[5] 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," in *Computer Speech and Language*, vol. 8, pp. 211-222, 1994.

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

[7] W. Goldenthal, and J. Glass, "Statistical Trajectory models for Phonetic Recognition," in *Procs ICSLP*, vol. 4, pp. 1871-1874, 1994.

[8] O. Ghitza, and M. Sondhi, "Hidden Markov models with templates as non-stationary states: an application to speech recognition," in *Computer Speech and Language*, 2, pp. 101-119, 1993.

[9] M. Ostendorf, and S. Roucos, "A Stochastic segment model (SSM) for phoneme-based continuous speech recognition," in *IEEE Transactions on ASSP*, vol. 37, pp. 1857-1869, 1989.

[10] J.-P. Martens, "A connectionist approach to continuous speech recognition," in *Procs FORWISS/CRIM ESPRIT Workshop*, pp. 26-33, Munich, 1994.

[11] K. Ponting, S. Peeling, "The use of Variable Frame Rate Analysis in Speech Recognition," Computer Speech and Language, vol. 5, no. 2, pp. 169-180, 1991.

[12] F. Jelinek, and R. Mercer, "Interpolated estimation of Markov source parameters from sparse data," in *Pattern Recogn. in Practice*, eds. Gelsema, E. and Kanal L. North Holland, Amsterdam, pp. 381-397, 1980.