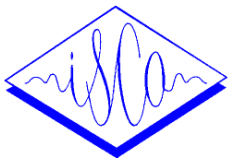

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

PHONE RECOGNITION USING A TRANSITION-CONTROLLED, SEGMENT-BASED DP/MLP HYBRID

Jan Verhasselt and Jean-Pierre Martens†*

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

ABSTRACT

In this paper, we explore how 'phone transition' models can be efficiently integrated into a segment-based continuous speech recognizer using Multi Layer Perceptrons. In this transition-controlled recognition strategy, the acoustic continuum of speech is modeled as a sequence of transitions between phones, but the advantages of segment-based modeling are retained by performing a kind of non-linear interpolation with context-independent phone probabilities. Experiments show that the transition-based phone recognizer significantly outperforms the context-independent one.

1 INTRODUCTION

In this paper, we describe how phone transition models can be integrated into a segment-based Dynamic Programming (DP) / Multi-layer Perceptron (MLP) system for continuous speech recognition [1]. These models are particularly suited for taking into account the coarticulation phenomena between consecutive phones. Several implementations are explored, each one representing a compromise between theoretic abilities, and practical considerations such as trainability and generalizational power. The simplest implementation consists of a single-output MLP, called the 'Transition Controlled Neural Net' (TCNN). As a first step, the performance of this TCNN-based system was evaluated on a phone recognition task.

It has been argued in many papers [1,2,3,4] that a segment-based approach to continuous speech recognition could offer definite advantages over the traditional frame-based approach. In particular, we mention the flexibility of introducing prior knowledge about speech (such as phone durations), and the capability of capturing the spectral/temporal relationships over the whole phone. The Stochastic Segment Model (SSM) [3] approach mostly uses a time-synchronous dynamic programming search to identify the best matching sequence of segment models for a given utterance. To be successful, many different segmentations have to be examined, leading to an extra loop [4] with respect to the Viterbi search in an HMM system. To be discriminative, the SSM system described in [1], uses MLP's to model the different units.

Although the aim of an acoustic phonetic decoding system 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: there are no distinct boundaries between the acoustic realizations of subsequent units. Instead, there are continuous transitions emerging from coarticulation phenomena. Recently, it has even been argued [5] that the identity of a phone is not uniquely determined by its short-time spectrum, but by temporal evolutions of this spectrum. Moreover, some phones, such as voiced stops [5] and nasals [6] are preferably described by formant transitions towards the phone and away from the phone. This implies that the acoustical realization of a phone is not only determined by the identity of the phone itself, but also by the identities of the preceding and the following phone (and to a smaller degree by the further context). These coarticulation phenomena are usually captured by context-dependent phone models. When introducing such models, one must choose between including more models to capture more coarticulation phenomena, and keeping the number of models restricted as to guarantee trainability and generalization to new tasks.

Alternatively, one could describe speech as a sequence of transitions between phones, and use MLP's to model these transitions. The acoustical observations needed for these phone transition models are found in the vicinity of the phone boundaries, and are mainly determined by the identity of the two adjacent phones (and to a smaller degree by the identity of the further context). Since these transition models are not modeling entire phones anymore, the transition-controlled system does not require the extra loop which is typical for a SSM recognizer. The number of probability estimates accumulated in the total score for a given utterance is equal to the number of potential phone boundaries that is being considered during the search. This number does not depend on the path being examined.

However, in order to preserve the advantages of a SSM system, we will supply the transition-modeling MLP's with phone probabilities emerging from a segment-based context-independent module, in addition to the inputs describing in detail the acoustical observations in the vicinity of the boundary being investigated.

A common approach to derive robust estimates for a large number of (context-dependent) probabilities is to interpolate them with the estimates of more general and better-trained (context-independent) models. The interpolation weights are typically determined by deleted interpolation [7]. By providing the transition-modeling MLP's with both transition specific inputs and context-independent phone probabilities, one may expect these MLP's to perform a kind of non-linear interpolation between the two types of inputs. The weights of this interpolation are learned during the Error Back Propagation (EBP) training of the nets.

*Supported by N.F.W.O. - IBM Grant
†Senior Research Associate of the National Fund for Scientific Research

2 THE BASELINE SYSTEM

The transition models are added to the baseline context-independent phone recognition system described in [1]. This system incorporates an auditory model front-end, an initial segmentation stage and a MLP-based phonetic classification and segmentation unit. The auditory model generates a sequence $\bar{o}$ of observation vectors each characterizing a 10 ms speech frame. The initial segmentation module generates a set $\bar{b}$ of N_b candidate phonetic segment boundaries. The segments of speech enclosed by two consecutive boundaries are called 'initial segments'. Candidate phonetic segments are built by concatenating up to four consecutive initial segments, and a MLP is trained to estimate the posterior probability that a given candidate phonetic segment is a true phonetic one. 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 major inputs of these MLP's are derived from the observation vectors describing the segment and its close surroundings. They are supplemented by segmental features such as the duration of the segment.

The DP-process examines several candidate phonetic segmentations $\bar{s}$ and phone sequences $\bar{u}$, and calls the MLP's to determine the probabilities of these phonetic decodings, given the acoustic evidence.

Thanks to the initial segmentation, the computational load can be kept low: for each new candidate phonetic boundary, only four preceding candidate boundaries need to be examined as possible starting points of a phonetic segment.

3 THE TRANSITION APPROACH

The transition models are designed to examine the spectral transitions at the boundary between two phones in order to identify one of these phones. Since those transitions are assumed to occur on the boundaries in $\bar{b}$, the phonetic decoding can be described as a sequence $\bar{t}$ of N_b transitions between phone-pairs, and $\bar{s}$ and $\bar{u}$ can be derived from $\bar{t}$. If the phone estimated at the previous boundary b_{l-1} was u_{l-1}, 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. 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. $u_{l-1} = u_l$).

The probabilistic framework and the implementational requirements are closely related. In the following paragraphs, we will introduce consecutive approximations in the probabilistic framework, in order to reduce the number of free parameters that have to be determined from the training corpus. This reduction is necessary in order to assure the trainability of the nets and the generalization to other tasks, given the limited amount of training data that was available. It is not our goal to present an exhaustive list of all possible implementations of the transition-based approach. We merely want to illustrate the reasons for the proposed compromises between the ability to learn specific characteristics and the trainability of the networks. The optimal implementation will depend on the size of the training corpus.

3.1 General Concepts

The posterior probability of $\bar{t}$, given the acoustical observations and the initial segmentation, can be factorized as follows:

$$P(\bar{t}|\bar{o},\bar{b}) = \prod_{l=1}^{N_b} P(t_{u_l}|t_{u_{l-1}}..t_{u_1}u_o,\bar{o},\bar{b}) \qquad (1)$$

Estimating these probabilities would require the training of MLP's for different lengths of the left-context. This would lead to a large number of models which are bound to learn specific characteristics of the corpus, given the limited number of examples for each context. Moreover, the search problem would require an unacceptably high computational effort.

If the phonetic context is constrained to the preceding phone, and if the acoustical observations and boundaries are represented by a small set of observations in the vicinity of the present boundary as well as by a set of context-independent phone probabilities emerging from the baseline SSM system, then the above expression can be approximated by:

$$P(\bar{t}) = \prod_{l=1}^{N_b} P(t_{u_l}|u_{l-1},\bar{v}_l(\bar{o},\bar{b}),\bar{p}_l(u_1),..,\bar{p}_l(u_K)) \qquad (2)$$

In this expression, $\bar{v}_l(\bar{o},\bar{b})$ represents a selection of observations in the vicinity of b_l, and $\bar{p}_l(u_j)$ represents context-independent probabilities of the phone u_j in some candidate phonetic segments in the vicinity of b_l. The formula in (2) represents a first order Markov process, which is extremely suitable for a left-to-right Viterbi search. Taking into account more context than u_{l-1} would lead to a higher order Markov process (and thus more dimensions in the search) and a larger number of models.

Although the transition models use phone probabilities derived for different candidate phonetic segments, the search performs a simultaneous segmentation and classification, thus eliminating the extra loop which is typical for the SSM approach. By only examining transitions t_{u_l} receiving sufficient evidence for u_l from the context-independent module, the CPU-time can be reduced drastically, without affecting the recognition performance.

A straightforward estimation of the posterior probabilities in (2) could be accomplished by means of K multi-layer perceptrons (one for each left context u_{l-1}), each having an output node for every right phone u_l. One of these neural networks is depicted in figure 1a. In this way, the weights of the connections between the inputs and the hidden layer are only shared among phone-pairs with the same left-phone. Such an implementation would require that a sufficient amount of examples of each phone-pair were available in the training corpus. As we do not have such a corpus at our disposition, we will try to restrict the number of free parameters in our system by sharing more neural network connections among the different phone-pairs.

Instead of designing one MLP for each left-context, one could for instance train a single MLP, with the identity of the left-context being explicitly supplied at the inputs of that MLP (like in [8]). This implementation is depicted in figure 1b. In this way, the total number of output nodes is reduced drastically, at the expense of a small increase in the number of inputs. The weights of the connections between the inputs and the hidden layer are now 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. One can consider the hidden units as building left-context controlled transformations of the observations and phone probabilities. These

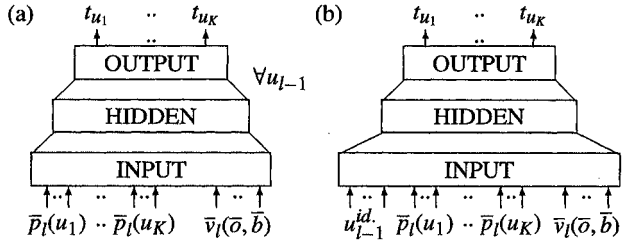

Figure 1: *Two different MLP implementations. The $u^{id.}_{l-1}$ inputs specify the identity of u_{l-1}.*

transformations will of course be much less phone-pair specific as the transformations that could be built in a system comprising one MLP per left-context, at least if one wants to restrict the number of hidden units.

An important problem with regard to two suggested implementations is that they require a large number of inputs. Especially the set of context-independent phone probabilities can be rather large: for each phone, several probabilities corresponding with different candidate phonetic segments have to be provided. In order to avoid this problem, one can create a single MLP with only one output, and use it in a hypothesis-testing scheme. On each boundary, the different possible transitions are hypothesised one at the time. The network estimates the posterior probability of the hypothesised transition. Since the same output is used for every transition, such a network must be supplied with the identity of u_l in addition to the observations and the identity of u_{l-1}. However, it would no longer be necessary to include in the inputs other context-independent phone probabilities then those computed for u_{l-1} and u_l (figure 2a). Consequently, one then obtains:

$$P(t_{u_l}|u_{l-1},\bar{v}_l(\bar{o},\bar{b}),\bar{p}_l(u_{l-1}),\bar{p}_l(u_l)) \qquad (3)$$

as approximations of the probabilities in equation (2). Note that in every node of the search space, the constructed MLP now needs to be called as many times as there are right-phones to investigate. Furthermore, this network still requires a substantial number of inputs for encoding the phone-pair (2 K inputs in case of one-of-n encoding).

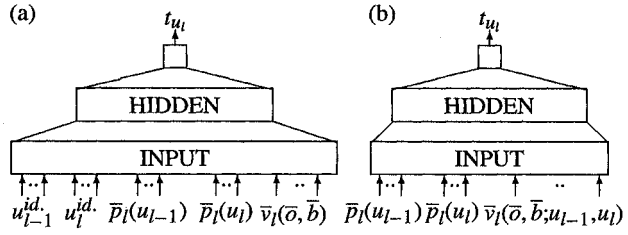

Figure 2: *(a) MLP-structure with explicit identification of the identity of the p hone-pair; (b) TCNN-structure.*

3.2 The TCNN Approach

To overcome the problem of large input vectors, we propose to replace the explicit identification of (u_{l-1},u_l) by an implicit one, and to train a MLP to estimate:

$$P(t_{u_l}|u_{l-1},\bar{v}_l(\bar{o},\bar{b};u_{l-1},u_l),\bar{p}_l(u_{l-1}),\bar{p}_l(u_l)) \qquad (4)$$

with $\bar{v}_l(\bar{o},\bar{b};u_{l-1},u_l)$ representing phone-pair specific transformations of the observations, giving an indication of the correspondence (or difference) between the actual acoustical observations and those which are typical for the investigated transition (figure 2b). Since these observations are typical for the transitions, we call this MLP a 'Transition Controlled Neural Net' (TCNN). Note that in this scheme, the MLP does no longer know the identity of the investigated transition and consequently it will not be able to learn the prior transition probabilities anymore. Furthermore, the success of this MLP will to a large extent depend on the phone-pair specific transformations $\bar{v}_l(\bar{o},\bar{b};u_{l-1},u_l)$ that are used to obtain the inputs. In fact, all features constructed by the MLP are now shared by all phone-pairs and everything that is really phone-pair specific has to be captured in the construction of the MLP-inputs.

One possibility for an appropriate $v_l(\bar{o},\bar{b};u_{l-1},u_l)$ would be to compute it with a MLP which is trained to estimate the posterior probabilities $P(t_{u_l}|u_{l-1},\bar{v}_l(\bar{o},\bar{b}))$ (e.g. implemented as the K-output MLP with explicit u_{l-1} encoding described above, but not taking into account the context-independent phone probabilities). Since the identity of the transition is uniquely specified to this MLP, it is able to learn the prior transition probabilities, so that these would be used implicitly in the MLP estimating (4). Note that in this case, the latter MLP actually performs a non-linear interpolation between the posterior *transition* and the context-independent *phone* probabilities. Since the identity of the transition is not specified to the MLP estimating (4), these interpolation weights are identical for all phone-pairs (which is not the case for weights determined by deleted interpolation).

3.3 Current Implementation

Instead of trying to obtain the posterior transition probability $P(t_{u_l}|u_{l-1},\bar{v}_l(\bar{o},\bar{b}))$ as an input, we have chosen to use difference measures quantifying the differences between the actual acoustical observations $(\bar{o},\bar{b})$ observed in the vicinity of b_l, and those which are typical for (u_{l-1},u_l). The TCNN is a MLP with one hidden layer, supplied with three sets of inputs. When examining a possible transition from u_{l-1} to u_l (given u_{l-1}), one set of inputs is composed of context-independent probabilities of u_{l-1} and u_l in candidate phonetic segments in the vicinity of the boundary being analyzed. A second set of inputs describes the durations of these segments in comparison with the expected duration of u_l, when succeeding u_{l-1}. Finally, there are also inputs representing the differences between the actual acoustical observations across the examined boundary and those which are typical for the transition from u_{l-1} to u_l. These acoustical observations are transformed to features describing movements in the spectral balance, and changes in the total energy and voicing evidence. Some features describe movements and changes in a confined region around the boundary, while others are less effected by the exact position of the boundary. In order to limit the number of free parameters, the typical feature values for a particular phone-pair (u_{l-1},u_l) are characterized by the maximum likelihood estimates of their means $\mu(u_{l-1},u_l)$ and their standard deviations $\sigma(u_{l-1},u_l)$. If x_m represents an acoustic feature, its difference measure which serves as an input to the TCNN is given by:

$$v_{lm}(\bar{o},\bar{b};u_{l-1},u_l) = \frac{|x_m - \mu_m(u_{l-1},u_l)|}{\sigma_m(u_{l-1},u_l)} \qquad (5)$$

No prior assumptions 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 EBP training of the TCNN to decide on how to combine the context-independent probability estimates and these context-dependent difference measures. The price we had to pay for the enormous reduction in free parameters is that this way of combining evidence is independent of the phone-pair. Furthermore, the transitions are examined in a rather primitive way.

Note that there are no nodes in this TCNN which are phone-pair specific, nor are there any inputs revealing the identity of the phone-pair being investigated, nor its probability. The TCNN is therefore not able to learn the prior probabilities of the phone-pairs, but only the prior probability of a correct hypothesis. However, the TCNN is able to use the prior probabilities of the phones, since these are implicitly provided with the posterior phone probabilities emerging from the context-independent module.

For every phone-pair having an insufficient number of occurrences in the training corpus to allow a reliable estimation of the means and variances, the average durations comprised in input set 2 are replaced by average durations of u_l disregarding u_{l-1}, and the difference measures of set 3 are replaced by default values (at present 0.67 for all variables). These inputs can also be provided for transitions between phone-pairs having no examples at all in the training corpus. These constitute the so called 'contextual holes'. Since in this way only the input sets 1 and 2 carry information for those transitions, the TCNN is forced to base its estimate on these inputs. This approach has the advantage that the TCNN learns how to handle contextual holes during its training, since phone-pairs with too few examples are treated as if they had no examples at all. In this way, it should be able to generalize to new tasks.

4 CORPUS AND TRAINING

The TCNN and the context-independent module were trained on a corpus 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 [1]. The training utterances were automatically labeled by aligning them to their phonetic transcriptions [9], and the means and the variances characteristic for the phone-pairs were computed from these alignments. For every candidate phonetic boundary b_l, input vectors for the TCNN were stored in a training database for each transition t_u (including the correct t_{u_l}) that needs to be examined. Each vector is assigned the label 'correct' or 'incorrect' to indicate whether there is really a transition from u_{l-1} to u. The TCNN is then trained using the EBP algorithm.

5 EXPERIMENTAL RESULTS

The TCNN-based system was evaluated on a phone recognition task. The performance was compared to that of the stochastic segment MLP/DP hybrid which also provided the context-independent phone probabilities. The test corpus consisted of 130 hand-labeled sentences originating from 10 speakers who were not in the training corpus. Since the TCNN in its current implementation is not able to learn the prior probabilities of the phone-pairs, the results should be compared with those of the baseline system using a unigram language model. As can be derived from table 1, the TCNN system yields a significant improvement over the baseline system. If the TCNN outputs are linearly interpolated with the prior phone-pair probabilities, an additional improvement is obtained. Note that the TCNN system continues to outperform the baseline system when the latter is using a bigram language model as well. The TCNN comprising 951 free parameters, was added to a baseline system with 8215 free parameters.

	Baseline system		TCNN	
	Unigram	Bigram	Unigram	Bigram
D	9.5%	12.3%	10.0%	12.6%
I	7.4%	4.0%	6.1%	4.7%
S	26.7%	25.3%	25.0%	23.1%
T	43.6%	41.6%	41.1%	40.4%

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

6 CONCLUSIONS

We have shown that a transition-controlled recognition strategy is an efficient way of introducing context-dependent modeling in a phone recognizer. Even when implemented as a simple TCNN, the new strategy yields a significant improvement over the context-independent system. This indicates the high potential of the new system when moving to the more sophisticated implementations outlined in section 3.

REFERENCES

[1] J.P. Martens, A. Vorstermans, N. Cremelie, "A New Dynamic Programming/Multi-Layer Perceptron Hybrid for Continuous Speech Recognition," EUROSPEECH, Berlin, Vo. 3, pp. 1937-1940, Sept. 1993.

[2] H. C. Leung, J. R. Glass, M. S. Phillips, V. W. Zue, "Detection and Classification of Phonemes Using Context-Independent Error Back-Propagation," ICSLP, Kobe, Vo. 2, pp. 1061-1064, Nov. 1990.

[3] M. Ostendorf, S. Roucos, "A Stochastic Segment Model (SSM) for Phoneme-Based Continuous Speech Recognition," ASSP, Vo. 37, pp. 1857-1869, 1989.

[4] R. M. Schwartz, Y. Chow, M. Dunham, O. Kimball, M. Krasner, F, Kubala, J. Makhoul, P. Price, S. Roucos, "Acoustic-Phonetic Decoding of Speech," in NATO ASI Series *Recent Advances in Speech Understanding and Dialog Systems*, Berlin Heidelberg: Springer-Verlag, Vo. F46, 1988.

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

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

[7] L. R. Bahl, F. Jelinek, R. L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, Vo. PAMI-5, pp. 179-190, 1983.

[8] N. Morgan, H. Bourlard, C. Wooters, P. Kohn, M. Cohen, "Phonetic Context in Hybrid HMM/MLP Continuous Speech Recognition," Eurospeech, Vo. 1, pp. 109-111, Sept. 1991.

[9] A. Vorstermans, J.P. Martens, "Automatic Labeling of Speech Synthesis Corpora ," ICSLP, Sept. 1994