THE IMPORTANCE OF SEGMENTATION PROBABILITY IN SEGMENT BASED SPEECH RECOGNIZERS

Jan Verhasselt[1]* *Irina Illina*[2] *Jean-Pierre Martens*[1] *Yifan Gong*[3] *Jean-Paul Haton*[2]

[1]ELIS, University of Ghent, St.-Pietersnieuwstraat 41, B-9000 Gent, Belgium
[2]CRIN/CNRS, INRIA-Lorraine BP239, 54506 Vandœvre-lès-Nancy, France
[3]Speech Research (PSL), Texas Instruments, P.O. 655303 MS 8374, Dallas, TX 75265, USA

ABSTRACT

In segment based recognizers, variable length speech segments are mapped to the basic speech units (phones, diphones,...). In this paper, we address the acoustical modeling of these basic units in the framework of segmental posterior distribution models (SPDM). The joint posterior probability of a unit sequence $\underline{u}$ and a segmentation $\underline{s}$, $Pr(\underline{u}, \underline{s}|\underline{X})$ can be written as the product of the segmentation probability $Pr(\underline{s}|\underline{X})$ and the unit classification probability $Pr(\underline{u}|\underline{s}, \underline{X})$, where $\underline{X}$ is the sequence of acoustic observation parameter vectors. In particular, we point out the role of the segmentation probability and demonstrate that it does improve the recognition accuracy. We present evidence for this in two different tasks (speaker dependent continuous word recognition in French and speaker independent phone recognition in American English) in combination with two different unit classification models.

1. SPD MODELING

An appropriate training algorithm for segmental posterior distribution models (SPDM) is an iterative algorithm that alternates between the following two steps:

1. Find the unit sequence $\hat{\underline{u}}$ and the corresponding segmentation $\hat{\underline{s}}$ satisfying:

$$\hat{\underline{s}}, \hat{\underline{u}} = \underset{\underline{s},\underline{u}}{\text{argmax}}\, Pr(\underline{u}, \underline{s}|\underline{X}, \theta)$$

   The model parameters are represented by θ and the maximization is over all possible segmentations $\underline{s}$ and over all admissible unit sequences $\underline{u}$ for the given "transcription" of the training data. In the remainder of this paper, we call $\hat{\underline{s}}$ the *unit segmentation*.

2. Find a new set of model parameters θ' yielding a higher joint probability estimate for the obtained unit and segment sequence:

$$Pr(\hat{\underline{u}}, \hat{\underline{s}}|\underline{X}, \theta') \geq Pr(\hat{\underline{u}}, \hat{\underline{s}}|\underline{X}, \theta)$$

As the maximization of $Pr(\hat{\underline{u}}, \hat{\underline{s}}|\underline{X}, \theta)$ necessarily implies a suppression of the $Pr(\underline{u}, \underline{s}|\underline{X}, \theta)$ with $(\underline{u}, \underline{s}) \neq (\hat{\underline{u}}, \hat{\underline{s}})$ (since $\sum_{\underline{s},\underline{u}} Pr(\underline{u}, \underline{s}|\underline{X}) = 1$), the parameter estimation algorithm (step 2) has to take all possible segmentations into account.

*Aspirant N.F.W.O. – Belgacom

Since the number of possible segmentations for an utterance of length F frames is 2^{F-1}, this is computationally expensive. In order to reduce the number of possible segmentations, two basic strategies are commonly used:

1. Use a preprocessing technique (e.g. frame clustering or segment boundary detection) [1, 2].
2. Use a non-segmental recognition system (e.g. an HMM) to provide a set of sentence hypotheses (an N-best list or a word-lattice) and restrict the segmentations to those observed in that set [3].

Although these methods reduce the number of possible non-unit segmentations considerably, they cannot remove them all without introducing errors. Since the parameter estimation algorithm of segmental *posterior* distribution models has to take the remaining non-unit segmentations into account, its training is computationally more expensive than the Viterbi training of segmental *likelihood* distribution models [4, 5]. However, in the next section we will show that introducing the segmentation probability permits a reduction of the SPDM training time.

2. SEGMENTATION PROBABILITY

2.1. Motivations For Using It

The joint posterior probability $Pr(\underline{s}, \underline{u}|\underline{X})$ can be written as the product of a *segmentation probability* and a *unit classification probability* [6]:

$$Pr(\underline{s}, \underline{u}|\underline{X}) = Pr(\underline{s}|\underline{X}).Pr(\underline{u}|\underline{s}, \underline{X}) \quad (1)$$

This factorization has several advantages:

- Dedicated statistical models can be trained for estimating the segmentation and the unit classification probabilities. Only the segmentation model has to be trained on all possible segmentations. It turns out however that an adequate segmentation model requires much less free parameters than an adequate unit classification model. The large unit classification model must only be trained on the unit segmentation since the latter occurs in the conditioning part. This yields a dramatical reduction of the training time.
- Since the unit classification models have to be trained on the unit segmentation only, it is possible to use the various re-estimation techniques that were originally developed for the training of segmental *likelihood*

distribution models [5]. The resulting likelihood estimates are then converted to posterior probabilities using Bayes' law:

$$Pr(\underline{u}|\underline{s},\underline{x}) = \frac{p(\underline{u},\underline{s},\underline{x})}{\sum_{\underline{v}} p(\underline{v},\underline{s},\underline{x})} = \frac{Pr(u)Pr(\underline{s}|\underline{u})p(\underline{x}|\underline{u},\underline{s})}{\sum_{\underline{v}} Pr(v)Pr(\underline{s}|\underline{v})p(\underline{x}|\underline{v},\underline{s})} \quad (2)$$

where the sum extends over all possible unit sequences.

- The computational requirements of the recognition process can be reduced considerably by calculating the unit probabilities only for those candidate segmentations having a segmentation probability exceeding a predefined threshold.

Omitting the segmentation probability in equation 1 would have severe consequences:

- It would boil down to assuming that all segmentations are equally probable. However, if information on the segmentation can be found in $\underline{x}$, then different candidate segmentations *will* have different posterior probabilities and the segmentation probability can have an important effect on the recognition result.
- If the unit classification models process segmental features, the segmentation probability addresses the conditioning event mismatch problem that is mentioned in [5]. A segmental feature vector is the result of a transformation $f(\underline{x}, s_{i-1}, s_i)$ of the parameter vector sequence $\underline{x}$ with the purpose of describing the variable length segment $]s_{i-1}, s_i]$ in its context.[1]. One common type of segmental feature vector is a fixed length, sampled version of the parameter vector sequence. The advantages of segmental over frame-based features are that they represent a broader (and possibly more interesting) range of feature extractions and that they allow to take full advantage of the correlation that exists among the parameter vectors in the segment (and its surroundings). Using segmental features, equation 1 becomes:

$$Pr(\underline{s},\underline{u}|\underline{x}) \equiv Pr(\underline{s}|\underline{x})Pr(\underline{u}|F(\underline{s},\underline{x})) \quad (3)$$

where $F(\underline{s},\underline{x})$ is the segmental feature vector sequence given $\underline{s}$, and '$\equiv$' indicates a modeling assumption. Since the segmental feature transformation depends on the segmentation, and since the recognition process has to evaluate different candidate segmentations $\underline{s}$, the overall conditioning event $F(\underline{s},\underline{x})$ is not unique. In this case, the foundation of statistical decision theory seems to be lost since the theory holds for comparing $Pr(u|z)$ to $Pr(u'|z)$ but not for the comparison of $Pr(u|z)$ to $Pr(u'|z')$. However, in our opinion, this potential problem is relaxed by the presence of the segmentation probability in equation 3. Indeed, suppose that we would have an ideal segmentation probability estimator at our disposal. This detector would generate a probability 1 for the correct segmentation and 0 for all other candidate segmentations. In this case, the unit probabilities would only have to be calculated for the segments of the correct segmentation, such that the overall conditioning event would be unique. Of course, in practice the segmentation probability estimator will not be ideal and the problem will not disappear completely. Whether this drawback of segmental features is more important than the intrinsic advantages of using them in the first place (see above) is an empirical question which is not yet fully answered.
- Even if no segmental or fixed-length feature vectors are used, modeling the segmentation probability is still motivated by the fact that the unit classification models are trained on unit segments only. It is very difficult to predict what probabilities will be produced for non-unit segments that were never seen during the training. If the unit models extrapolate appropriately, they will produce low unit probabilities for these "unseen patterns", and the segmentation probability will not add much to the recognition performance. If, on the other hand the unit models extrapolate badly on non-unit segments, then the multiplication by the segmentation probability in equation 1 will help to suppress the unit probabilities on the non-unit segments.

We found evidence in the literature that the suppression of unit probabilities on non-unit segments *is* important in segmental posterior distribution model based recognition. In the Segmental Neural Net (SNN) approach [3], the segmentation probability is not incorporated explicitly, but a significant improvement of the recognition performance is obtained by training the SNN negatively on segments belonging to incorrect hypotheses in an N-best list and differing from the segments found in the correct hypothesis (we would call this the non-unit segments in the N-best list). As a result, the SNN is trained to produce low outputs (i.e. low unit probability estimates) on the non-unit segments. In an earlier implementation (which was called "1-best training"), the SNN was exclusively trained on the unit segmentation, i.e. it was used to estimate the unit classification probability $Pr(\underline{u}|\underline{s},\underline{x})$. The N-best algorithm on the other hand, trains the joint posterior probability $Pr(\underline{u},\underline{s}|\underline{x})$, (i.e. the left hand side of equation 1) using the N-best list for reducing the number of possible segmentations. The fact that the N-best algorithm was found to be significantly better than the 1-best algorithm, indicates that the segmentation probability *is* important.

In section 5.3., we present additional experimental evidence for the importance of the "unseen pattern effect".

2.2. Modeling The Segmentation Probability

If we define s_i as the index of the final frame of the i^{th} segment corresponding with the i^{th} basic unit u_i, the segmentation model can be factorized and approximated by:

$$Pr(\underline{s}|\underline{x}) = \prod_{i=1}^{L(\underline{u})} Pr(s_i|s_0,...,s_{i-1},\underline{x}) \approx \prod_{i=1}^{L(\underline{u})} Pr(s_i|s_{i-1},\mathbf{Y}_i) \quad (4)$$

[1] Note that we use a broader definition than [5]: we allow parameter vectors from outside the segment in the transformation, which enables us to relax certain conditional independence assumptions

with $L(\underline{u})$ being the length of $\underline{u}$, and $\mathbf{Y}_i$ being a fixed length segmental feature vector describing the variable length segment $]s_{i-1}, s_i]$ in its context.[2] The probability $Pr(s_i|s_{i-1}, \mathbf{Y}_i)$ is called the *segment probability*. Original work in this area [6] proposed to calculate this probability as a product of boundary probabilities, estimated by a Multi-Layer Perceptron (MLP). In our segment modeling approach however [2], we also compute the boundary probabilities but just as intermediate results in the segment probability calculation. We estimate the segment probability by means of a so-called segment-MLP with one output. It is trained on all segments that can be hypothesized in the recognition process and that start on a true unit boundary. The teaching output is 1 for segments belonging to the unit segmentation $\underline{s}$, and 0 for all others. The least mean squares cost function was minimised with the standard backpropagation algorithm. The inputs consisted of the duration d_i of the candidate segment (i.e. $s_i - s_{i-1}$) and of $\mathbf{Y}_i$. The vector $\mathbf{Y}_i$ was composed of three types of segmental features:

- Change functions measuring changes in the parameter vectors $\mathbf{x}$ in the segment and in the vicinity of its boundaries. These functions are the first order derivatives of the parameters, the Spectral Variation Function [7] and the correlation between successive vectors.
- The final segment boundary probability and the maximal segment-internal boundary probability.
- Averages of the parameter vectors in the initial and the final part of the segment (the partition is determined by the location of the segment-internal boundary with maximal boundary probability).

3. UNIT CLASSIFICATION PROBABILITY

The unit classification probability is often approximated by context-independent models. One then obtains:

$$Pr(\underline{u}|\underline{s}, \underline{x}) \approx \prod_{i=1}^{L(\underline{u})} Pr(u_i|s_i, s_{i-1}, \mathbf{Z}_i) \qquad (5)$$

with $\mathbf{Z}_i$ being, in our current systems, a fixed length segmental feature vector representing the acoustic observations observed in the i^{th} segment. The experiments described in this paper were performed using two different recognisers incorporating different ways of modeling the *unit probability* $Pr(u_i|s_i, s_{i-1}, \mathbf{Z}_i)$: the Stochastic Trajectory Model (MSTM) recogniser developed at CRIN [8], and the Discriminative Stochastic Segment Model (DSSM) recognizer developed at ELIS [2].

3.1. MSTM Unit Probability Computation

In the MSTM recognizer[8], $\mathbf{Z}_i$ is a fixed length vector that is composed of Q parameter vectors. These Q vectors are derived from the variable length sequence of parameter vectors $\mathbf{x}$ observed within the segment $]s_{i-1}, s_i]$. The likelihood of $\mathbf{Z}_i$, given the duration d_i and the identity of the unit u_i is modeled as:

$$p(\mathbf{Z}_i|d_i, u_i) \triangleq \sum_{k \in T_{u_i}} p(\mathbf{Z}_i|t_k, d_i, u_i) Pr(t_k|d_i, u_i) \qquad (6)$$

[2]It is not required that $\mathbf{Y}_i$ is a fixed length segmental feature vector, it just happens to be the case in our current systems

spk	alv	dof	flf	loc	ols	pab	yfg	AVG
MSTM	1.62	4.12	0.94	3.92	1.15	1.96	2.02	2.49
MSTMS	0.94	3.85	1.01	0.74	0.74	1.01	1.89	1.45

Table 1. Total word error rates [%] (per speaker and the mean over the speakers) using the MSTM recognizer in speaker dependent mode, without (MSTM) and with (MSTMS) segmentation probability

where T_{u_i} is the set of component trajectories associated with u_i. Using equation 2, the posterior probability of the basic unit u_i is computed as:

$$Pr(u_i|s_i, s_{i-1}, \mathbf{Z}_i) = \frac{p(\mathbf{Z}_i|d_i, u_i) Pr(d_i|u_i) Pr(u_i)}{\sum_v p(\mathbf{Z}_i|d_i, v) Pr(d_i|v) Pr(v)} \qquad (7)$$

where the summation extends over the complete set of basic units.

3.2. DSSM Unit Probability Computation

In the DSSM recognizer [2], a Multi Layer Perceptron (called the *unit MLP*) is trained to estimate the unit probabilities $Pr(u_i|s_i, s_{i-1}, \mathbf{Z}_i)$. The input vector of the MLP consists of the duration d_i and the acoustical evidence $\mathbf{Z}_i$ obtained from the parameter vectors $\mathbf{x}$ observed in the segment $]s_{i-1} - 2, s_i + 2]$. The sigmoidal MLP output nodes correspond to the basic units and were trained by error backpropagation, using a least mean squares cost function. The teaching outputs were 1 for the correct unit and 0 for all the other units.

4. EXPERIMENTS AND RESULTS

4.1. MSTM Recognizer

The experiments with the MSTM recognizer deal with the French corpus CEA recorded by the CRIN laboratory. The acoustic parameter vectors are 13^{th} order mel-cepstral vectors including a normalized energy. Context-independent models are built for 32 phones, including one silence model. The language model has a word-pair equivalent perplexity of 48 and a vocabulary of 2010 words. Table 1 shows the total error rates (including deletion, insertion and substitution errors) for speaker dependent continuous word recognition. The table contains the results for two situations: MSTM (using the standard formulation without segmentation probability) and MSTMS (MSTM with segmentation probability). We observe that adding the segmentation probability reduces the mean error rate by 42%. This is a statistically significant improvement (the 95% confidence interval for MSTM is 2.2% – 2.8% and for MSTMS it is 1.2% – 1.7%).

4.2. DSSM Recognizer

The experiments with the DSSM recognizer deal with acoustic-phonetic decoding on the American English corpus TIMIT. Each acoustic parameter vector $\mathbf{x}$ consist of an auditory spectrum, a voicing evidence and a total energy. The reported results, are results for the 39 phones set defined in [9]. Context-independent phone models were trained for these 39 phones, and for the glottal stop. During the recognition experiments, we used a phone bigram

# par's	23K	56K	146K	205K	295K
DSSMnoS	41.79	40.42	39.34	36.69	36.62
DSSM	36.53	35.25	34.07	33.37	32.96

Table 2. Total phone error rates [%] of five DSSM recognizers without (DSSMnoS) and with (DSSM) segmentation probability.

(for these 40 phones) language model. For the evaluation, the glottal stop was removed from both the recognized and the true phone sequence. Training was performed on the sx and si sentences of the NIST designated training corpus. Testing was performed on the sx and si sentences of the complete NIST designated test corpus. We have trained five systems, with different recognition performances. The recognizers differ in the size of the unit MLP, but they all use the same segmentation model. Table 2 shows the total phone recognition error rates (TE) (deletion + insertion + substitution errors) as a function of the number of parameters in the unit MLP.

Again we observe that adding the segmentation probability gives a statistically significant improvement of the recognition (the 95% confidence interval for the best DSSM recognizer is 32.55% – 33.37%, and without segmentation probability it is 36.20% – 37.04%).

4.3. Relation To The "Unseen Pattern Effect"

In order to correlate the importance of the segmentation probability to the importance of the "unseen pattern effect" mentioned in section 2.1., we have analyzed the posterior unit probability estimates on unit and non-unit segments. The analysis was carried out for the five DSSM recognizers that were described in the previous section. For each of the recognizers, we have calculated the average of the maximal unit probability estimate over all non-unit segments. This average is denoted AMPNU. Similarly, AMPU is the average maximal unit probability estimate over all unit segments. First of all, we observed that AMPU is always larger than AMPNU, indicating that the MLP's have the tendency to consider the non-unit segments as corresponding to none of the units. However, this tendency is not very firm. Secondly, we noticed that the relative reduction of the total phone error rate (RER) due to introducing the segmentation probability

$$RER = \frac{TE(DSSMnoS) - TE(DSSM)}{TE(DSSMnoS)}$$

is strongly correlated with AMPNU (correlation coefficient = 0.765, see figure 1). We can thus conclude that the less successfully the unit classification model suppresses the unit probabilities on non-unit segments, the more important the segmentation probability (which has the same suppressing effect) becomes.

5. CONCLUSION

In this paper, we have pointed out the role of the segmentation probability [6] in segment-based recognition systems incorporating segmental posterior distribution models. A significant improvement of the word recognition accuracies

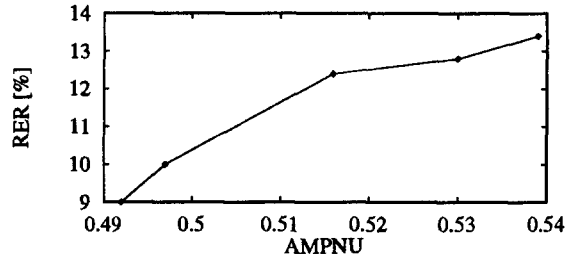

Figure 1. Relative reduction of the phone error rate (RER) vs. average maximal unit probability estimate on non-unit segments (AMPNU).

was obtained by adding the segmentation probability to the MSTM [8] word recognizer. Experiments with the DSSM [2] recognizer confirmed that the phone recognition accuracy too is significantly improved by incorporating the segmentation probability. Additional experiments showed that the importance of the segmentation probability is strongly correlated with the inability of unit classification models to suppress the unit probabilities on non-unit segments.

6. ACKNOWLEDGMENT

This work originates from a cooperation within COST action 249: continuous speech recognition over the telephone.

REFERENCES

[1] V. Zue, J. Glass, M. Philips, and S. Seneff, "Acoustic Segmentation and Phonetic Classification in the SUMMIT system," *Proc. of ICASSP89*, Vol. 1, pp. 389-392.

[2] J.-P. Martens, "A Connectionist Approach to Continuous Speech Recognition," *Procs FORWISS/CRIM ESPRIT Workshop*, pp. 26-33, Munich, 1994.

[3] S. Austin, G. Zavaliagkos, J. Makhoul, and R. Schwartz, "Speech Recognition using Segmental Neural Nets," *Procs ICASSP*, Vol. I, pp. 625-628, 1992.

[4] M. Ostendorf, and S. Roucos, "A Stochastic Segment Model (SSM) for Phoneme-based Continuous Speech Recognition," *IEEE Trans. on ASSP*, Vol. 37, pp. 1857-1869, 1989.

[5] 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.

[6] H. Leung, J. Glass, M. Phillips and V. Zue, "Detection and Classification of Phonemes Using Context-Independent Error Back-Propagation," *Procs ICSLP*, pp. 1061-1064, 1990.

[7] G. Flammia, P. Dalsgaard, O. Andersen, and B. Lindberg, "Segment Based Variable Frame Rate Speech Analysis and Recognition Using a Spectral Variation Function," *Procs ICSLP*, pp. 983-986, 1992.

[8] Y. Gong and J.-P. Haton, "Stochastic Trajectory Modeling and Sentence Searching for Continuous Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 5, No. 1, 1997.

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