INTEGRATION OF CONTEXT-DEPENDENT DURATIONAL KNOWLEDGE INTO *HMM*-BASED SPEECH RECOGNITION

Xue Wang, Louis F. M. ten Bosch & Louis C. W. Pols*

Institute of Phonetic Sciences / IFOTT, University of Amsterdam,
Herengracht 338, nl-1016 CG Amsterdam, the Netherlands, e-mail: wang@fon.let.uva.nl
*Lernout & Hauspie Speech Products N.V., Brussels, Belgium

ABSTRACT

This paper presents research on integrating context-dependent durational knowledge into HMM-based speech recognition. The first part of the paper presents work on obtaining relations between the parameters of the context-free HMMs and their durational behaviour, in preparation for the context-dependent durational modelling presented in the second part. Duration integration is realised via rescoring in the post-processing step of our *N*-best monophone recogniser. We use the multi-speaker TIMIT database for our analyses.

1. INTRODUCTION

It is commonly known that the duration information is not well integrated into the standard HMM-based automatic speech recognition (ASR) techniques. A major repair of the durational behaviour of HMM has been the use of hidden semi-Markov model (HSMM) with explicit state durational pdf's [2]. However such a technique has not been used in most state-of-the-art ASR systems, due to its computational complexity [4]. On the other hand, simply implementating the minimal duration of the standard HMM is being used repeatedly without knowing the exact behaviour of such HMMs. Such a situation causes difficulties in further attempting to integrate durational knowledge into HMM-based ASR in a systematic way. Integrating other long-term features such as pitch, will cause similar problems.

The present paper reports about our experiences of conscious integration of durational knowledge into ASR based on standard HMM (non-HSMM). In Sect. 2 analyses will be given of the durational behaviour of the HMM and it will be revealed that even the simplest linear HMMs provide rich possibilities in modelling the phonetic segmental duration. Theoretical relations will be given between the HMM parameters and the durational pdf (dpdf) that the HMM governs. In Sect. 3 we will report about ways to integrate the constraints of segmental duration into the Baum-Welch ML training procedure and we will give results of improving the recognition and automatic segmentation performance using monophone HMMs. In Sect. 4 analyses of context-dependent duration statistics, using the TIMIT database, will be briefly discussed. In Sect. 5, proposals and an actual implementation will be presented to integrate the context-dependent duration models into a post-processing process of our monophone-based recogniser. In the final section, future research based on the experience obtained in this work is pointed out.

2. DPDF OF STANDARD *HMM*

The single-state dpdf (which is geometrical) is less important than the dpdf of the whole HMM, because in actual practice it is the latter that models a phonetic segment. In this section we firstly derive the closed-form whole-model dpdf for general left-to-right HMM. Left-to-right is by far the most common type of transition topology used for speech recognition. It may include any number of skipping transitions and parallel paths but no feed-back loops that contains more than one state. Then an analysis will be given of the general properties of the dpdf, with the help of some examples of useful topologies.

2.1. Obtaining the dpdf of the whole HMM

First we analyse the dpdf of the simplest *linear* HMM which is a cascade of states each having a selfloop. From probability theory [5], this dpdf is the convolution of all the dpdf's of the states in the cascade, since the total time in the HMM, as a random variable (r.v.), is the sum of the independent r.v.'s of the time in the states. Using z-transform for help, the major part (for convenience) of the dpdf of a cascade with a total of n states is

$$\hat{P}_n(z) = \prod_{k=1}^{K} \frac{1}{(z-a_k)^{n_k}} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \frac{C_k^i}{(z-a_k)^i}. \qquad (1)$$

where some subsets of n_k states may have the same selfloop probabilities a_k (the locations of these equal selfloops can be arbitrary). The coefficients C_k^i in the right-hand side can be obtained by partial fraction decomposition. Each summation term has a simple dpdf form in the d-domain, i.e. a negative binomial:

$$\binom{d-1}{i-1} a_k^{d-i} v(d-i) \Leftrightarrow \frac{1}{(z-a_k)^i}, \qquad i = 1,2,\cdots,n_k,$$

where v is a step function and $\Leftrightarrow$ denotes z-transformation. One can then inversely z-transform the terms in (1) back to the d-domain and then sum them to get the dpdf. In general such dpdf has a complicated form, especially for large number K of different subsets. For instance, for a cascade of 10 states in 4 subsets of 1, 2, 3 and 4 states with equal selfloop probabilities, respectively, the lowest-order binomial has 54 terms in the numerator of its coefficient. So we will not print it here. The general dpdf is

$$P_n(d) = \hat{P}_n(d) \prod_{k=1}^{K} (1-a_k)^{n_k}.$$

Any general left-to-right topology (including skips) can be easily decomposed into separate linear paths. Then the total dpdf is simply a combination of the dpdf's of all the paths. The analysis given above for the closed-form dpdf gives some insight into its function form. For numerical calculation purposes, however, one can use a property of the Markov chain to get the dpdf by multiplying the transition matrix to itself for *d* times.

2.2. Analysis of whole-model dpdf

It turned out that HMMs, even with complicated topologies, have a single-peak, binomial-like dpdf, except for models with *independent paths*. Independent paths are defined here as linear paths that do not share any state with selfloop from another path. Below we present examples (with numerical values) to reveal some of the properties of the dpdf. In Fig 1, the K.F. Lee model [1] shows three separate bars in the dpdf (two are high) due to the three independent lower paths (without selfloops). The second model shows a single-peak dpdf because it does not contain any independent path.

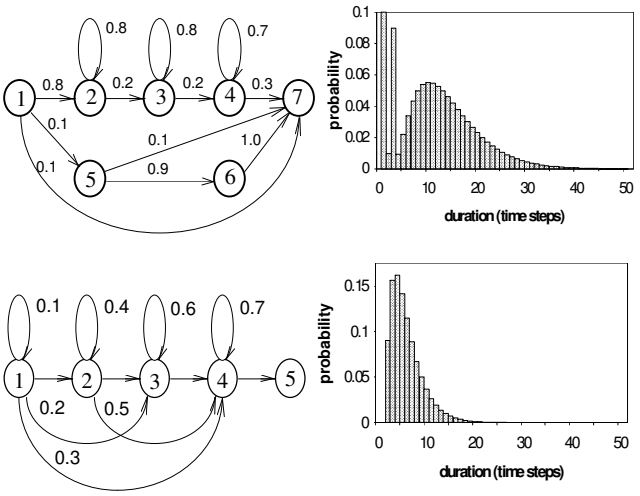

Figure 1: In the upper part the K.F. Lee model is shown and belowe a left-to-right model with skips. The corresponding dpdf's are shown on the right-hand side. The time steps are general, and equal to the frame-shift of the recogniser (8 ms in our case).

The following formulae show the relations between the duration mean μ, variance σ^2, selfloop probability a_i and length *n* of linear models (single path), which are also shown in Fig 2 for the even simpler case of all-equal selfloop probability (the right-most equalities in the formula). It can be seen that even such a simple linear model provides rich possibilities in modelling segmental duration, usually having a binomial-like shape. Proper values of the parameters should be chosen in order to get the actual fit, since the parameters still allow for a lot of freedom.

$$\mu = \sum_{i=1}^{n} \frac{1}{1-a_i} = \frac{n}{1-a}; \quad \sigma^2 = \sum_{i=1}^{n} \frac{a_i}{(1-a_i)^2} = \frac{na}{(1-a)^2}.$$

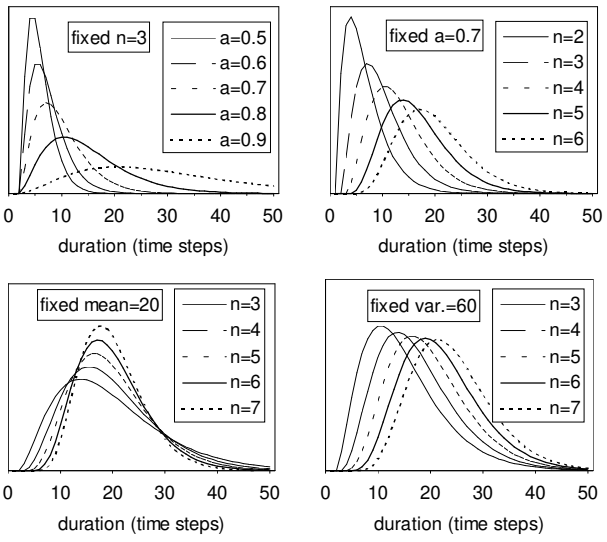

Figure 2: Durational behaviour of linear HMMs upon varying single parameters. Vertical axes are for probability values.

3. *ML*-TRAINING CONSTRAINED WITH DURATIONAL STATISTICS

It appears that the dpdf modelled with the HMM trained with the usual ML generally does not fit the data statistics very well . This is because the usual Baum-Welch ML training procedure does not take global information at the segment level into account. Our approach is to integrate the durational statistics of the phone segments into the ML training procedure. First the duration μ and σ^2 of all the phones in the system are collected and they together define the possible *n* for each monophone HMM [7]:

$$\frac{\mu(\mu-1)+\sigma^2}{\mu-1+\sigma^2} < n < \mu+1-\sqrt{2\sigma^2+1}.$$

Then the ML equations with duration statistics μ and σ^2 as extra constraints are

$$\begin{cases} \dfrac{D(i,i)}{\tilde{a}_i} - \dfrac{D(i,i+1)}{1-\tilde{a}_i} + \theta_1 \dfrac{1}{(1-\tilde{a}_i)^2} + \theta_2 \dfrac{1+\tilde{a}_i}{(1-\tilde{a}_i)^3} = 0, \quad i = 1,2,\cdots n; \\ \displaystyle\sum_{i=1}^{n} \frac{1}{1-\tilde{a}_i} = \mu; \\ \displaystyle\sum_{i=1}^{n} \frac{\tilde{a}_i}{(1-\tilde{a}_i)^2} = \sigma^2, \end{cases}$$

where *D* are the 'counts' obtained from one previous ML iteration (the same as in the usual ML), and θ's are two La grange multipliers. It turned out that this set of *n*+2 equations cannot be solved analytically as in the usual case. Newton-Raphson procedures [3] were used to solve this set of non-linear equations together with well chosen initial points for search. For some phones, the data μ and σ^2 need to be modified. The fitting situation is shown for two example phones in Table 1. Durational

statistics are collected from the whole TIMIT data (for further details about this constrained training see [7]). The TIMIT training set was used to train the 50 monophone HMMs with 3 Gaussians per state, a diagonal covariance matrix, and *n* chosen as above. Table 2 shows word-correct score improvements in both recognition and segmentation.

phone	original data		modified			modelled		modelled + c	
	μ	σ	μ	σ	*n*	μ	σ	μ	σ
y	8.34	4.40		3.49	4	14.94	6.48	8.34	3.49
z	10.51	3.91			5	12.59	4.56	10.51	3.91

Table 1: Original data μ and σ, for the phones /y/ and /z/ For semi-vowel /y/ μ and σ are modified, this is also shown. *n* is the chosen model length. The two right-most columns show μ and σ modelled by the usual HMM and by HMM trained with durational constraints (+ c). Units for μ and σ are in 8 ms steps (frame shift of the recogniser).

	no constraint	with duration constraint
recognition	80.61% (77.73%)	86.83% (84.41%)
segmentation	83.48%	84.48%

Table 2: Word correct and accuracy (in brackets) scores and segmentation accuracy with a 20 ms margin on both directions.

4. ANALYSIS OF CONTEXT-DEPENDENT DURATIONAL STATISTICS

Based on the statistical analysis on duration distribution using 11 contextual factors [6], 4 factors were actually chosen for context-dependent modelling for recognition, as shown in Table 3. For each of the 1323 cells in the factorial design, the duration μ and σ were calculated from the TIMIT training `sx` and `si` utterances. Then parametrical models were made for each cell using the binomial-like dpdf's (see section 2), because of their well suited shapes for phone duration. In calculating these dpdf's, Markov models were used (different from the phone HMMs) [7]. All cells with one or more observations were fitted. The dpdf's for empty cells (unseen data) were left all-zero, indicating the impossibility of those combinations of factor levels in the training data.

factor:		levels
R	speaking rate	0: fast 1: average 2: slow
S	stress (of vowels)	0:not 1:primary 2: second.
Lw	syllable location in word	if *S*=0,1: 0:rest; 1:final; 2:penul. 3:mono if *S*=2: 0:rest; 1:final; 2:penul.
Lu	syl. location in utterance	if *Lw*=0: 0:rest if *Lw*=1: 0:rest;1:final; 2:penul. if *Lw*=2: 0:rest; 2:penul. if *Lw*=3: 0:rest;1:final; 2:penul.

Table 3: Contextual factors and their levels (penul=penultimate).

5. INTEGRATION OF *CD*-DURATION MODELS IN POST-PROCESSING

The durational knowledge, stored in the 1323 context-dependent (CD) duration models, was integrated in the recogniser in a re-scoring process based on the *N*-best transcriptions. An *N*-best program providing transcriptions at both the word and the phone level was not available to us. So, in order to re-score the transcriptions at the word level, using our CD duration models at the phone level, a two-phase procedure was developed (Fig 3). In the first phase the *N*-best word transcriptions were generated and in the second phase the phone-level transcriptions were generated based on the lexical form of each word and an optional word-juncture model.

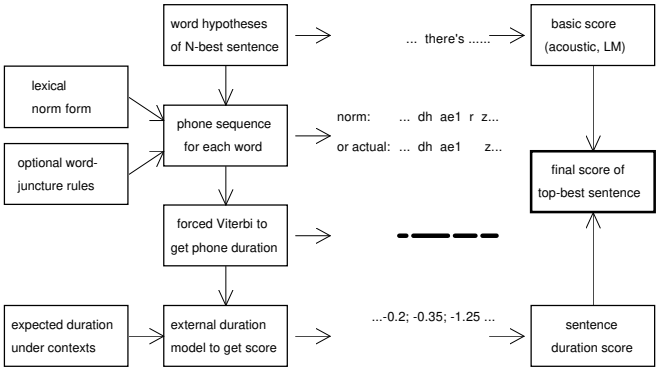

Figure 3: Procedures of re-scoring *N*-best transcriptions using CD duration models. Illustrated in the middle (from top to bottom) are, a transcribed word, the norm and an 'actual' phone transcription derived from the word-juncture model, the estimated phone duration, and the duration scores of the phones.

5.1. Word-juncture modelling

Modifications from norm phone to actual phone realisation were only performed at the word junctures. This is based on the observation that most non-norm phone transcriptions occur at word borders [6]. The word juncture model was derived from the `sx` and `si` training utterances, and is given as a list of rules. The input of each rule is a sequence of phones in the *juncture region* of two adjacent words, as given in the lexicon for the two words. The output is the actual phone sequence according to the most-frequent realisation in the training set. The juncture region here includes, either all consonents, or a single vowel, of both words. Below is an example, where "." indicates the word border.

input		output
cl k cl t.cl t	⇒	cl t

5.2. Duration score

The duration scores from the CD dpdf's for phones have to be combined into an utterance-level duration score. Direct summation (in logarithm) would emphasise the effect of the difference in number of phones in each utterance transcription from the *N*-best output. We used two procedures to normalise for

this effect. The first normalisation is at the phone level and uses four typical values of duration defined on a phone dpdf s_d (Fig. 4). These are the duration for maximum point $d_i^{\max}$ of s_d for phone i, the CD duration mean μ_i, the duration normalised d_{i0} over the utterance transcription, and the actual duration d_i. Two differences are used for relative duration shifts:

$$\Delta_i^A = \begin{cases} -\left|s_d(d_{i0}) - s_d(\mu_i)\right|, & \text{if } d_{i0} \geq d_i^{\max}; \\ s_d(\mu_i) + s_d(d_{i0}) - 2s_d(d_i^{\max}), & \text{otherwise}; \end{cases}$$

$$\Delta_i^B = \begin{cases} -\left|s_d(d_i) - s_d(d_{i0})\right|, & \text{if } (d_i - d_i^{\max})(d_{i0} - d_i^{\max}) > 0; \\ s_d(d_{i0}) + s_d(d_i) - 2s_d(d_i^{\max}), & \text{otherwise}. \end{cases}$$

The second normalisation occurs at the utterance level taking into account the total number of phones *I*. The duration score for an utterance is the (weighted) sum of the two difference terms:

$$s_D = \frac{1}{I}\sum_{i=1}^{I}(w_l\Delta_i^A + \Delta_i^B).$$

The total score for an utterance is the weighted sum (with another weighting factor w_d) of s_D and the score of the utterance obtained in the *N*-best recognition process.

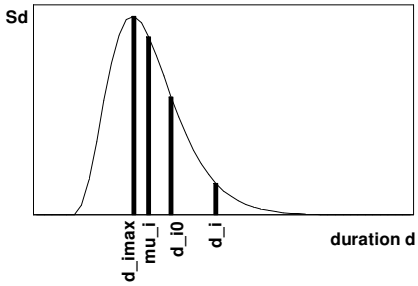

Figure 4: Four typical duration values for a dpdf.

5.3. Re-scoring

The baseline system for this test had 50 monophone HMMs, each with the number of states *n* determined as in section 3 (ranging from 3 to 10). However the observation pdf's of these states were further tied to 3 for each phone. Each such pdf had 8 Gaussian components, each with a diagonal covariance matrix. In the *N*-best process only 655 utterances, of the total of 1344 `sx` and `si` utterances in the TIMIT test set, had errors in the top best transcriptions. We applied the two-phase duration modelling procedure in this section only to this set of "wrong" utterances. Top 20 transcriptions were generated with a word-lattice *N*-best algorithm and were used in the re-scoring process. Due to the inaccuracy introduced by the separation of the phone-level transcription from the word-level transcription, only a very tiny increment in word correct score (3 more words correct than without re-scoring) was obtained on this "wrong set" with a well chosen weighting factor w_d. Experiments on the whole set of 1344 utteranes were not yet performed, but it has to be expected that the word-correct score for the whole set will only decrease after durational re-scoring with the current algorithm.

6. DISCUSSION

In this paper we tried to perform context-dependent duration modelling with context-free HMMs. The durational behaviour of the standard (linear) HMM is revealed to be rich (thus providing more than just a minimum duration), but the model parameters need to be adjusted to fit the data durational statistics at the segment level. Constrained training was performed and improvements were obtained. Context-dependent duration information was collected based on statistical analyses, and was then added to the system with a post-processing process using *N*-best transcriptions. Actual improvement in word-correct score was hindered by the lack of a good *N*-best algorithm. Future research will involve the implementation of our context-dependent duration models with a better *N*-best algorithm, and a systematic optimisation among the various system components.

The experience in this work shows a possibility to integrate long-term speech features into the frame-level-based HMM technique. Integrating context-dependent information (of e.g. duration or pitch) in the last step of post-processing provides a simple system structure, and has the advantage of being able to correct any modelling errors that might have been made by the statistical modelling of other system components, no matter how "perfect" these components are designed.

7. REFERENCES

1. Lee, K.-F. *Automatic speech recognition: the development of the Sphinx system,* Kluwer Academic Publishers, Boston, 1989.

2. Levinson, S. E., "Continuous variable duration hidden Markov models for automatic speech recognition", Computer Speech and Language*,* 1(1), 1986, pp 29-45.

3. Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. *Numerical recipes in Pascal,* Cambridge Univ. Press, 1989.

4. Mitcell C., Harper, M. & Jamieson, L "On the complexity of explicit duration HMM's", IEEE Transactions on Speech and Audio Processing, 3(3), 1995, pp 213-217.

5. Papoulis, A. *Probability & statistics,* Prentice-Hall, Inc., Englewood Cliffs, NJ, 1990.

6. Pols, L.C.W., Wang, X. & ten Bosch, L.F.M. "Modelling of phone duration (using the TIMIT database) and its potential benefit for ASR", presented to Speech Comm.

7. Wang, X. *Integrating knowledge on segmental duration in HMM-based continuous speech recognition*, Ph.D. thesis, University of Amsterdam, in preparation.