Phone duration models for fast broadcast news transcriptions

Benjamin Lecouteux, Driss Matrouf, Pascal Nocera, Georges Linares

PHONE DURATION MODELS FOR FAST BROADCAST NEWS TRANSCRIPTIONS

G. Linares, B. Lecouteux, D. Matrouf, P. Nocera

LIA/CNRS, Université d'Avignon
Agroparc, BP 1228
84911 Avignon CEDEX 9, France
georges.linares,benjamin.lecouteux,driss.matrouf,pascal.nocera@lia.univ-avignon.fr

ABSTRACT

Phone duration modeling in HMM based LVCSR systems has been largely studied during last few years. In this paper, we address the problem of duration modeling in the particular context of fast decoding on LVCSR task. Discrete distributions are integrated in the LIA's broadcast news transcription system, and influence of duration modeling is studied using various pruning schemes. Experimental results show that duration modeling improve significantly the pruning efficiency.
In a second time, we show that durations are intrinsecly acoustic-context dependent. Crossed experiments are conducted combining context independent acoustic models and context dependent duration models. We show how durations are affected by acoustic context.
At last, we propose a rate dependent modeling of phone durations. This method outperforms significantly our rate independent duration model based system. Globally, integration of rate dependent models allows a absolute WER gain of 3.3%.

1. INTRODUCTION

HMM based systems are known to be deficient in modeling phone durations. The implicit duration distribution of Markov Models is determined by state to state transitions. Therefore, durations follow a geometric distribution which do not reflect the true duration distributions. This observation has been the start point of several works proposing and evaluating various modeling techniques. Most authors propose to plug explicit duration functions into HMMs, using probability density functions, such as gaussian, gamma functions ([1]), or discrete distributions ([2]). Other works propose approaches based on low level signal analysis ([3]), or specific HMM topologies which model durations in a more realistic way([4]).
Most of these studies have been achieved on limited tasks (digit or letter recogniton, small vocabulary,...), on LVCSR using large recognition systems, where the acoustic space is deeply explorated. In a context of fast decoding process -less than 5x Real Time (RT)- the search space is reduced to the few best hypothesis. We suggest that introduction of explicit duration models in a fast decoding engine may reduce the rank of best hypothesis in the decoding stack, leading to improvement of both decoding speed and recognition performances. We investigate this idea using the LIA's Broadcast News System. Experiments are achivied on the French radiophonic database developed for the evaluation campaign ESTER ([5]). First, we integrate discrete duration models to the recognition system. Two pruning schemes are defined, corresponding to 2xRT and 10xRT systems on a 3Ghz *pentium 4* processor. Results with and without duration modeling are compared, and the influence of pruning on duration efficiency is evaluated.
In a second part, we study the influence of acoustic context on duration modeling. Context dependent duration models and context independent ones are evaluated using various acoustic units.
An other point largely studied concerns the effect of speech rates on phone durations ([6]). In the last section, we describe and evaluate a method for rate-dependent models estimation, and their integration to the full recognition process.

2. INTEGRATING PHONE DURATION MODELS IN LVCSR SYSTEM

2.1. LIA's Broadcast News System

2.1.1. ESTER framework

Experiments are carried out on the ESTER-2003 French corpus. ESTER is an evaluation campaign of rich text transcription systems. This project is organized by AFCP (*Association Française de Communication Parlée*), the French departement of defense. ESTER-2003 is the first part of ESTER corpus, which has been provided for dry run tests. It is made up of 40 hours of transcripted radiophonic shows which are collected from two radio stations: *France Inter* (25h) and RFI (*Radio France International*, 15h). Shows

contain various acoustic conditions (large/wide band, noisy/-clean, musical background, etc.) and various speech modes (broadcast news, interviews, read or spontaneous speech, etc.). Moreover, linguistic resources includes a text corpus extracted from 4 years of the French newspaper *Le Monde*. Here, experiments are conducted using the two shows from the full ESTER corpus where our baseline system obtains the best and the worse recognition rates (France Inter 1998-17-12, 7h/8h show and RFI 2000-9-12, 9h30/10h30 show).

2.1.2. Transcription System

The full transcription system is composed of 3 independent steps.
First, speech signal is parametrized by 39 coefficient vectors: 12-mel cepstral coefficients plus energy and their first and second order derivative parameters.
The second step performs a segmentation into acoustic macro-classes. The aim of this step is to split each show signal into segments which will be suitable for the recognition engine. This segmentation is achieved using a hierarchical GMM classifier. A first classification process extracts speech segments from the full shows. A second one splits segments into *clean-background* and *musical-background* parts. At last, a gender dependent segmentation is achieved. At the end of this step, cepstral mean subtraction and variance normalization are performed on each extracted segment. This process splits a show into 6 classes which will be processed by speech recognizer using specific acoustic models (gender and background dependent).
Third step consists in the decoding process. The LIA's speech recognition system is based on classical HMMs for acoustic modeling and n-gram language models. The next section describes the specific search algorithm of our recognition engine and the associated pruning strategy.
For ESTER decoding, we use a 20k word lexicon extracted from the train corpus. Language model results of a linear interpolation of an ESTER specific trigram model and a journalistic model trained on the *Le Monde-1987-2002* database.

2.1.3. Speeral

Experiments are conducted using SPEERAL [7], the LVCSR system developed at the LIA. This system is able to use n-gram language models and classical 3-state left-to-right HMMs for acoustic modeling. Context dependent models can be used including cross word contexts.
Search engine is an asynchronous stack decoder based on a modified A* algorithm. Rather than word by word, the A* used in Speeral is based on a phoneme lattice previously generated.
The search complexity is reduced using several pruning strategies. First, beam size and extensive backtracks are controlled by static thresholds. Moreover, a cache memory stores partially explorated paths, saving new exploration of already evaluated paths. At last, a fast probe function allows a good selection of well scored paths, satisfying A* specific constraint of probe function optimality. A full description of Speeral can be found in [7].
Using full context acoustic models and large trigram language models, light pruning schemes lead to a decoding time of about 60x Real Time. This ratio can be reduced to 2xRT using only right diphones, fast likelihood computation algorithm and a very strict pruning scheme. Of course, such a velocity improvement leads to a significant WER increase, about 60% relative.

2.2. Integrating duration models

As explained in the last section, the LIA's recognition engine explores a phone graph. For efficiency reasons, this graphis is not really built in a first pass, but it is locally built and deleted during search. The cost of each developed branch is evaluated addressing request to acoustic models about likelihood of an observation sequence knowing a given acoustic unit and a fixed duration. Therefore, integration of duration probabilities into this process can be achivied easily, combining strictly acoustic scores to duration likelihoods. On the other hand, there is no information about the word crossed on each node and usage of word-dependent duration units should by very complex. Considering complexity and computational cost of integration of such information to this search algorithm, we have chosen to use phone-level durations, in spite of few reported experiments which suggest better performances obtained using word specific durations.

Discrete distributions are estimated collecting phone duration from the full train corpus, grouping all acoustic macro-classes from the first segmentation process. Therefore, phone lattices are built estimating the log-likelihood $LogL(x|M_i, T_i)$ of a frame sequence x of length $|x|$, knowing the acoustic model M_i and the duration model T_i:

$$LogL(x|M_i, T_i) = LogL(x|M_i) + \lambda.LogL(|x||T_i)$$

The first term represents the acoustic log-likelihood. It is estimated using a classical Viterbi decoding, without any temporal information. The computation of the second term (duration log-likelihood) consists in accessing directly to the histogram modeling duration of unit i. The duration fudge factor λ has been chosen empirically ($\lambda = 2$). We can see that additional computational cost of this method can be neglected, compared to the full decoding process cost.

3. PRUNING SCHEMES AND DURATION MODELS

The fist part of ESTER corpus includes shows from two French radios, France Inter and Radio France International

(RFI). Shows from RFI are significantly badly decoded by the 60xRT system. These lower performances are due to larger acoustic and linguistic variability, due to larger inter-speaker variability. We defined two pruning configuration of Speeral, allowing about 10xRT (F1 system) decoding and (F0 system) 2xRT decoding. These systems use diphone models WER obtained by these baseline systems are compared to systems using duration models (cf. table 1).

System	Word Error Rate (%)	
	F-Inter	RFI
F0	40.1	61.1
F0-DUR	39.2	56.6
F1	30.3	50.7
F1-DUR	30.4	48.3

Table 1:*Baseline 2xRT (noted F0) and 10xRT systems (noted F1 versus Duration models based systems (F0-DUR and F1-DUR)*

Results shows first that using the *slow* system, duration models lead to an absolute WER reduction of 2.4% on RFI show, and remains equivalent on FranceInter show (-0.1%). This evaluation confirms that duration models are more efficient on adverse acoustic conditions. Nevertheless, duration allows a significant improvement of fast system performances: absolute gain of 3.5% and 1.6 % are obtained respectively on RFI and FranceInter shows. These results confirm that duration models reduce the deep of the best hypthesis in the search lattice, specialy in bad acoustic conditions.

4. CONTEXT-DEPENDENT VERSUS CONTEXT INDEPENDENT DURATION MODELS

Likelihood computation consumes a major part of CPU time in fast ASR systems. In spite of fast likelihood computation techniques, this complexity depends significativly of HMMs complexity. We tested the influence of duration models on low complexity models, and we compared them to recognition rates obtained using our context dependent models. 38 acoustic units are modeled by 3-state HMMs. Each GMM is a mixture of 32 gaussian functions; therefore, these models provides very low complexity and allows real time ratio lower than 1xRT. Results of these experiments are reported in the table 2.

System	Word Error Rate (%)	
	F-Inter	RFI
F0-BASE	49.7	65.9
F0-CI-DUR	48.7	63.6
F1-BASE	41.6	59.4
F1-DUR	43.6	59.4

Table 2:*WER of baseline system based on context independent units (F2-BASE), using context independent durations (F2-CI-DUR) , and context dependent durations (it F2-CD-DUR). These last configuration outperforms (F2-CD-DUR); this result illustrate context dependence of phone durations.*

We can show that context independent models are not able to exploit temporal information contained by duration models in the F1 case: usage of phone duration affects system performances; WER don't change on RFI show but they increase of about 2% of FranceInter show. Nevertheless, light gains are still obtained using F0 system.

This experiment suggest that duration are intrinscely context dependent. In order to confirm this assumption, we perform an other experiment where context dependent duration models are plugged on context independent acoustic models. Only duration modeling differs between the two model sets. Results (cf. table 3) shows that usage of acoustic context duration models leads to WER decreasing, in spite of identical acoustic model precision. Another interesting point is that gain obtained on RFI and FranceInter shows are similar, in spite of best efficiency reported on RFI, using context dependent models. The weakness of such low complexity models seems to be partially compensated by temporal information.

System	Word Error Rate (%)	
	F-Inter	RFI
F2-BASE	49.7	65.9
F2-CI-DUR	48.7	63.6
F2-CD-DUR	45.9	62.1

Table 2:*WER of baseline system based on context independent units (F2-BASE), using context independent durations (F2-CI-DUR) , and context dependent durations (it F2-CD-DUR). These last configuration outperforms (F2-CD-DUR); this result illustrate context dependence of phone durations.*

5. RATE DEPENDENT PHONE DURATION MODELS

Generally, duration based systems do not take account of speech rate. Nevertheless, phone durations depends of the global rate of a speech segment. Moreover, deformations are phone dependent, and a simple linear normalization should not provides a good modeling of duration elasticity. [1] propose to define 3 rate schemes (low, medium and hight rate). For each of them, rate-dependent models are learned; recognition process estimate the speech rate a select dynamically the best rate dependent model. We adopt this approach defining 6 rates schemes. Each of them correspond to a speech rate band. These bands are determined in two step; first, the train corpus is splitted into 50 subset following the rate factor of each speech segment. During second step,

these set are merged into 6 subband. Rate Subband overlap themselves in order to get enough learn data for model estimation. So, rate dependent models are learned on each subset. The figure 1 shows rate repartition and chosen subands.

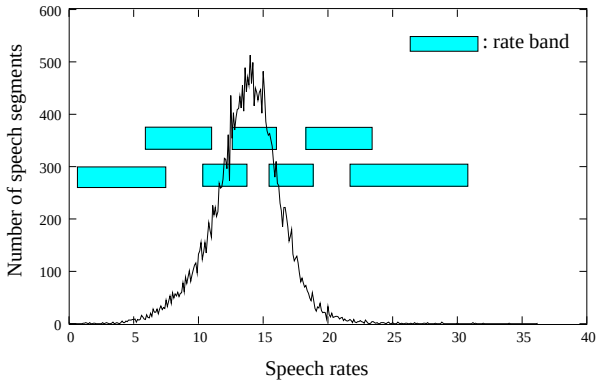

Fig. 1. *Speech rate repartion on ESTER corpus and rate subband. This curve shows the number of speech segment for for each speech rate ratio. Speech rate is expressed in number of frame per phone. The 6 subband correspond to 6 overlapping corpus subsets. Each of them is used for the learning of a rate dependent duration model.*

Before each recognition run, we estimate the rate performing a fast decoding. Usage of full recognizer instead of a simple phone recognizer is expensive in term of CPU time but allows more robust estimation of speech rate. This model selection process take about 1xRT. The second pass performs the decoding based on the duration model selected during the first pass. The table 4 shows results obtained using this technique.

	Word Error Rate (%)	
System	F-Inter	RFI
F0-BASE	40.1	61.1
F0-DUR	39.2	57.7
F0-RD	38.1	56.6

Table 4:*WER using Rate Dependent duration models (RD) versus standard durations models and 2xRT baseline system (F0-BASE).*

Results shows that rate dependent models outperforms classical ones of 1.1% (absolute) both on RFI and FranceInter shows, in spite of greater WER gain obtained on that last test signal using standard duration models. Cumulative WER absolute gains reach 2% and 4.5% on that two test shows.

6. CONCLUSION

We have studied phone duration modeling for fast decoding of broadcast news. Our experiments on a task of French broadcast news transcriptions show that usage of phone duration models can improve system performances. WER decreases significantly using restrictive pruning schemes, in spite of relatively low gains obtained using slower systems. Moreover, reported results using duration suggest that temporal information is more efficient when acoustic models fail.

At last, speech rate dependent models improve significantly performances of duration based systems. Nevertheless, this last method is based on a speech estimation which is cputime consuming.

We work now on the integration of a fast speech rate estimator in the search using the A* probe function. Moreover, we work now on estimation of duration models depending on the high level context, such as speaker or environment.

7. REFERENCES

[1] S.E. Levinson, "Continuously variable duration hidden markov models for automatic speech recognition," *Computer Speech and Langage*, vol. 1, pp. 29–45, 1986.

[2] C.D. Mitchell and L.H. Lamieson, "Modeling duration in a hidden markov model with the exponential family," in *International Conference on Automatic Speech and Signal Processing (ICASSP)*, Minneapolis (USA), 1993, pp. 331–334.

[3] J.P. Nedel and R. Stern, "Duration normalization for improved recognition of spontaneous and read speech via missing feature methods," in *International Conference on Automatic Speech and Signal Processing (ICASSP)*, Salt Lake City (USA), 1991.

[4] H. Ney, "Acoustic modelling of phoneme units for continuous speech recognition," in *Proc. EUSIPCO*, Barcelona, 1990, pp. 62–72.

[5] E. Geoffroy S. Galliano K. Mc Tait G. Gravier, J.F. Bonastre and K. Choukri, "Ester, une campagne d'évaluation des systèmes d'indexation d'émissions radiophoniques en français," in *Journées d'Études sur la Parole (J)*, Fes (Maroc), Mai 2004.

[6] D. Guiliani F. Brugnara, R. De Mori and M. Omologo, "A family of parallel hidden markov models," in *International Conference on Automatic Speech and Signal Processing (ICASSP)*, San Fanscisco (USA), May 1992, pp. 377–380.

[7] D. Massonié P. Nocera, G. Linares and L. Lefort, "Phoneme lattice based a* search algorithm for speech recognition," in *Text, Specch and Dialogue (TSD)*, Brno (Czech Republic), September 2002.