LATTICE RESCORING EXPERIMENTS WITH DURATION MODELS

Nicolas Jennequin and Jean-Luc Gauvain

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{jennequi,gauvain}@limsi.fr

ABSTRACT

This paper reports on experiments using phone and word duration models to improve speech recognition accuracy. The duration information is integrated into state-of-the-art large vocabulary speech recognition systems by rescoring word lattices that include phone-level segmentations. Experimental results are given for a conversational telephone speech (CTS) task in French and for the TC-Star EPPS transcription task in Spanish and English. An absolute word error rate reduction of about 0.5% is observed for the CTS task, and smaller but consistent gains are observed for the EPPS task.

1. INTRODUCTION

It is well known that HMMs do not properly model phone and word durations. Even if just the state duration is considered, regular HMMs do not offer a realistic model for duration [1]. The transition probabilities usually have no impact on the recognizer accuracy. It is often said that a uniform distribution can be more appropriate than the distribution given by the transition probabilities estimated on the training data.

When using triphone HMMs (with derivative features), the segment durations (state and phone) are encoded in the model topology and the derivative features in addition to the transition probabilities. None of these model parameters can properly capture segment duration when considering a context wider than a triphone. More specific duration models must be used to adequately model longer span durations. Duration features can be added at various levels of the acoustic models so as to represent HMM state durations, phone durations, and word durations. In this work only the phone and word durations are considered.

The duration models are used in a post-precessing step by rescoring word lattices with phone segmentations. The word lattices must include the phone segmentation for each word edge, i.e. the word lattices can be seen as phone lattices with lexical constraints. Lattice rescoring is a more desirable method than N-best rescoring for at least two reasons. First, rescoring word lattices instead of N-best lists is more accurate and more efficient. Second, N-best rescoring does not fit very well with consensus decoding which is known to significantly reduce the word error rate over a regular MAP decoding [6][1], i.e. the gain due to the use of duration models may be lost by backing off to MAP decoding if N-best rescoring is used.

In the following sections, the duration models used in this work are described along with how they are used in the speech recognizer. Experimental results are given for a conversational telephone speech (CTS) task in French and for the TC-Star European Parliament Plenary Sessions (EPPS) transcription task in Spanish and English.

2. DURATION MODELING

As stated in the introduction, HMMs do not properly model the speech segment durations, this is particularly true when only considering the transition probabilities. Figures 1 and 2 show the empirical distributions of the phone duration in the French CTS training data obtained after performing a forced alignment (via Viterbi decoding) of the orthographic transcriptions (with a dictionary allowing alternative pronunciations) and the acoustic data. The vowel distributions are given in Figure 1 and the consonant distributions are in Figure 2. Other phones not represented in these two figures are the semi-vowels and the special phones used for hesitations and pauses. It can be seen that these distributions are mostly unimodal. Some distributions are strictly decreasing (the schwa vowel and the liquid l) showing that the minimal duration imposed by the three state left-to-right HMM modeling each phone may not be appropriate for these phones. Looking at these distributions, it is apparent that phone duration can be helpful to differentiate the phones. The three state left-to-right topology implies that the pdf of the phone duration is the convolution of three geometric distributions. It is known that this does not reflect reality as is illustrated in Figure 3 where the empirical distribution for the triphone `e(s,t)` (phone /e/ in the context (/s/,/t/)) observed in the French CTS corpus is represented along with a 3-state geometric pdf, a gamma pdf, and a Gaussian mixture pdf whose parameters are estimated on the CTS training data. By comparing these distributions, it is clear that the Gaussian mixture pdf is a

[1] N-best consensus decoding is less effective than regular consensus decoding.

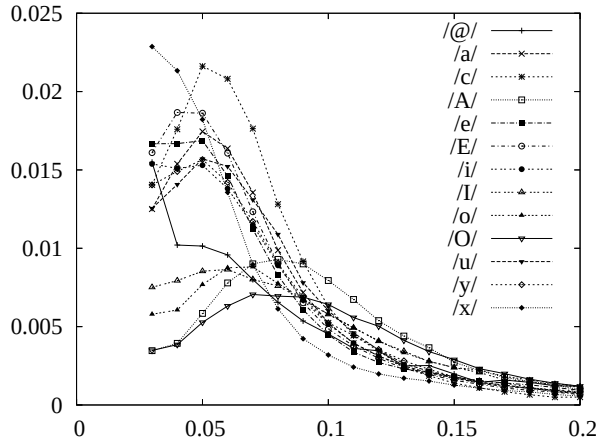

Figure 1: Empirical duration distribution for 13 vowels in the 140 hour French CTS corpus. (Horizontal axis in seconds.)

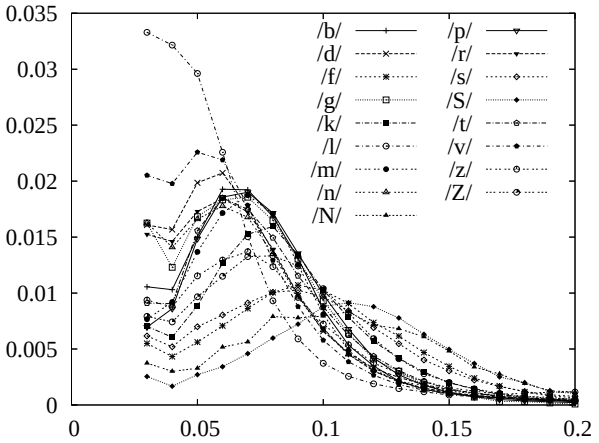

Figure 2: Empirical duration distribution for 17 consonants in the 140 hour French CTS corpus. (Horizontal axis in seconds.)

better approximation than the 3-state geometric pdf and the gamma pdf. The gamma pdf is shown here since it has often been used to model HMM state durations and phone durations [1, 5, 2, 8].

Directly modeling the word duration with a pdf [9] may be a viable solution for small vocabulary tasks, in particular for short words (1 or 2 syllables), but for very large vocabularies it is more appropriate to use a model with a back-off mechanism (in case a word rarely or never occurs in the training data), and with the capability to also model phone durations within a word. For this work, the model proposed by Gadde [7] was adopted, where each word is represented by a vector composed of the durations of the individual phones in the word. Phone and word durations are modeled with Gaussian mixtures, using word duration (seen as a vector of phones) when enough data is available to properly estimate the word model, and backing off to phone durations if this is not the case. As in [7] the duration models are used in a post-decoding step, but instead of applying such post-processing to an N-best list, it is applied to a word lattice. The augmented edge likelihood is the product of the HMM likelihood and the duration likelihood properly scaled.

3. MODEL ESTIMATION

Given a training corpus with orthographic transcriptions, the phone and word durations are obtained after forced alignment between the phone transcriptions (as given by the pronunciation dictionary) and the speech signal, using a set of tied-state context-dependent phone models. For all the experiments reported in this paper, the acoustic models include about 10K tied HMM states with 32 Gaussians per state (cf. the decoding section (Section 4 for more details about the recognizer models).

Given the phone segmentations, for each word pronunciation $H = (h_1, \ldots, h_{N_H})$ observed in the training data, the parameters of an N_H dimensional Gaussian mixture (GMM) representing the pdf $f(d_1, \ldots, d_{N_H} \mid H, W)$

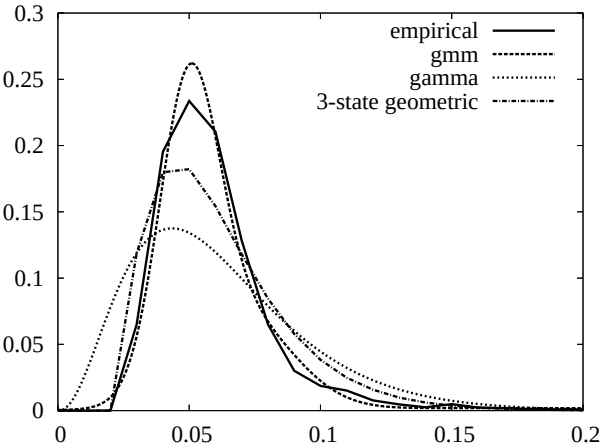

Figure 3: Duration pdf for the phone /e/ in context /(s,t)/ in the French CTS training corpus: empirical distribution; multigaussian distribution (4 Gaussians); gamma distribution; and convolution of three geometrical distributions (for a 3-state HMM). (Horizontal axis in seconds.)

are estimated, where $d_1 \ldots d_{N_H}$ are the duration of the phones in the word pronunciation H of the word W. As is usually done for GMMs, the pdf parameters are estimated by using the EM algorithm starting with a single Gaussian and iteratively splitting each Gaussian until the desired number of mixture components is reached. Since a very large vocabulary system is being targeted, the sparse training problem is a major issue as a large proportion of the words in the recognizer vocabulary are never or rarely observed in the acoustic training data. Table 1 contains the proportions of pronunciations with no more than n occurrences in the French CTS training data. The recognizer vocabulary contains 50K words and about 74K pronunciations. It can be seen that about 56% of the prononunciations[2] are never observed in the training data, and about 80% of the prononunciations occur at most twice, and only 8% of the pronunciations occur more than 10 times in the acoustic training data. This

[2]The term pronunciation here is used to refer to a particular pronunciation of a given word.

n	%pron	%occur
0	55.6%	0.9%
1	73.3%	1.4%
2	80.4%	1.9%
5	88.3%	2.9%
10	92.5%	4.1%

Table 1: Proportions of the vocabulary pronunciations with no more than n occurrences in the training data (French CTS data). The third columns gives the corresponding proportions of running words in the development data.

clearly demonstrates the data sparseness issue and the need for a smoothing and/or a back-off mechanism. The third column of Table 1 gives the corresponding proportions of the running words in the recognizer hypotheses on a set of development data (i.e. the counts are weighted by the word frequencies). These proportions show that words that are rare in the training data are generally also going to be rare in the data to be processed by the systems. Therefore, even though there is a data sparseness problem, it should not have too large an effect on word error rate since only 4% of the hypothesized pronunciations have a frequency count lower than 10 in the training data.

Combining MAP smoothing [3] and back-off to phone models [7] gave the best results on the development data. The prior pdf for each pronunciation model is obtained from the single Gaussian model of each phone composing the pronunciation. If the number of occurrences for a given pronunciation in the training data is lower than 20, we just back-off to the phone models. Here it should be noted that the back-off phone models can be GMMs whereas the MAP smoothing can only rely on a single Gaussian phone pdf, as there is no easy way to get an adequate prior pdf for a word pronunciation from the GMM phone models, even though diagonal covariances are used for the pronunciation duration vectors.

The multivariate word duration pdfs are therefore estimated as follow:

$$f(d_1,\ldots,d_{N_H}|H,W) = \begin{cases} f^m(d_1,\ldots,d_{N_H}|H,W) & \text{if } C(H,W)) > C_t, \\ \prod_{i=1}^{N_H} f(d_i|p_{H,i}) & \text{otherwise} \end{cases} \quad (1)$$

where $f^m(\cdot|H,W)$ is the MAP estimate of the duration model for the word pronunciation (H,W), $C(\cdot)$ is the frequency of the word pronunciation in the training corpus, $p_{H,i}$ is the i-th phone of the pronunciation H, and C_t is the frequency count threshold. In the current implementation both the MAP prior pdfs and the phone back-off pdfs are context independent. The threshold parameter and the pdf prior weight were optimized by maximizing the likelihood of the development data.

4. DECODING WITH DURATION MODELS

For the three systems on which experimental results are reported, decoding is carried out in multiple passes where the hypothesis of one pass is used by the next pass for acoustic model adaptation. For each decoding pass, the acoustic models are first adapted using both the CMLLR and MLLR adaptation methods. MLLR adaptation relies on a tree organization of the tied states to create the regression classes as a function of the available data. Then a word lattice is produced for each speech segment using a dynamic network decoder with a 2-gram or a 3-gram language model. This word lattice is rescored with a 4-gram language model and converted into a confusion network [6] taking into account the pronunciation probabilities. The words with the highest posterior in each confusion set are hypothesized along with their posterior probabilities. For the CTS data, the first hypothesis is also used to estimate VTLN warp factors for each conversation side [4].

The acoustic training data for the three systems (CTS French, EPPS English, and EPPS Spanish) include respectively 140h, 72h, and 79h of speech. The acoustic models used in the last decoding pass of each system include about 10K tied states with about 32 Gaussians per state. The respective vocabularies include 50K words, 60K words, and 65K words, with respectively 74k, 74k, and 94k pronunciations. The language models include 23M 3-grams and 15M 4-grams for French CTS, 33M 3-grams and 24M 4-grams of English EPPS, and 22M 3-grams and 45M 4-grams for Spanish EPPS.

During the last decoding step, a word lattice including the phone segmentation for each word edge is generated for each test segment. The duration log-likelihood as given in Equation 1 is then added to each edge log-likelihood score assuming that the acoustic models and the duration models are modeling independent variables. Two additional parameters are used to optimize the combination: a duration model weight to scale the word duration likelihood, and an additive constant proportional to the number of phones in the given word pronunciation. These two parameters have been optimized on the development test set for each task, but they are in fact pretty much task independent as the results are basically the same when the same values are used for the three tasks. After adding the duration scores to the lattice, the recognizer hypothesis is obtained by carrying out a consensus decoding in the same way it is done without the duration model.

An issue often raised about duration models is that word duration depend on the speaker, or more specifically, on the rate of speech. Therefore it may be desirable to normalize the word and phone durations by the rate of speech. This can be done by normalizing the overall average phone duration of data for each speaker in the training data and in the test data. Doing this normalization on the training data is easy since the phone durations can be scaled after forced alignment with the manual transcriptions. For the test data the lattice posterior probabilities

Conditions	DEV 2h	EVAL 2h	DEV+EVAL 4h
4-gram MAP	32.95	34.84	33.88
+ duration models	32.47	34.42	33.44
Consensus	32.03	33.77	32.90
+ duration models	31.61	33.32	32.45

Table 2: Word error rates on the French CTS development and test sets (28 conversation sides for each set, about 2h) using four rescoring configurations: 4-gram MAP decoding, MAP decoding with duration models, consensus decoding, and consensus decoding with duration models.

have been used to estimate the expected average duration given the current best models.

5. EXPERIMENTAL RESULTS

A first set of experiments has been carried out on the French CTS data. The best results are obtained using 2 Gaussians for the word duration models and 4 Gaussians for the phone models with a back-off threshold of 20 (the number of occurrences of the given word pronunciation in the training data). The HMM and duration models were trained on all the acoustic training data (140h) and the system parameters were optimized on the 2h development data set. The decoding parameters for the baseline system (i.e. without a duration model) have been carefully optimized and give word error rates of 32.0% and 33.8% respectively on the development data and the evaluation test set. The main results on this data are reported in Table 2. The use of duration models reduces the word error by about 0.5% absolute on the evaluation data (from 33.8% to 33.3%). It can also be seen that this gain is less than the gain obtained by consensus decoding compared to a standard MAP decoding. This shows the interest of using a decoding scheme for the duration models which is compatible with consensus decoding.

Experiments have also been carried out after normalizing the rate of speech in test and/or in the training data as described in Section 4, but these experiments resulted in no additional gain.

Results for the English and Spanish EPPS tasks are reported in Table 3. The data used for these experiments are the TC-Star Dev06 and Eval06 test sets. The Spanish data sets are about twice as large as the English sets as they also include about 3h of the Spanish Parliament data in addition to the EPPS data. These results are similar to those obtained on the French CTS data, but the error reductions are smaller since the baseline results are significantly better. It can also be seen that the duration models help the Spanish system more than the English system.

6. CONCLUSIONS

In the paper experiments have been reported showing that word and phone duration models can help reduce the word error rate of carefully optimized state-of-the-art

Conditions	English		Spanish	
	DEV 3.2h	EVAL 3.2h	DEV 6.1h	EVAL 7.0h
4-gram MAP	11.51	9.53	7.83	11.20
+ duration models	11.22	9.19	7.62	10.97
Consensus	10.84	9.05	7.55	10.85
+ duration models	10.71	8.86	7.39	10.61

Table 3: Word error rates on the English and Spanish EPPS development and test sets using four rescoring configurations: 4-gram MAP decoding, MAP decoding with duration models, consensus decoding, and consensus decoding with duration models.

LVCSR systems. The proposed approach is based on the rescoring of word lattices including phone segmentations for each word edge, thereby allowing the duration models to be compatible with a consensus network decoding framework. Experimental results were given for a conversational telephone speech task in French and for the TC-Star EPPS transcription task in Spanish and English. A word error rate reduction of about 0.5% absolute is reported for the CTS task, and smaller but consistent gains are reported for the EPPS task.

REFERENCES

[1] M. Russell and R.K. Moore, "Explicit modelling of state occupancy in hidden markov models for automatic speech recognition," *Proc. of IEEE Conference on Acoustics Speech and Signal Processing*, pp. 5–8, June 1985.

[2] D Burshtein, "Robust parametric modeling of durations in hidden markov models," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Detroit, pp. I-548–551, May, 1995.

[3] J.-L. Gauvain and C.H. Lee. "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. on Speech and Audio Processing*, 2(2):291–298, April 1994.

[4] J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen, and F. Lefevre, "Conversational Telephone Speech Recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, April 2003, pp. I–212–215.

[5] S.E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Language*, 1(1):29–45, 1986.

[6] L. Mangu, E. Brill, and A. Stolke, "Finding Consensus Among Words: Lattice-Based Word Error Minimization," in *ISCA Eurospeech*, Budapest, Sept. 1999, pp. 495–498.

[7] V.R.R. Gadde, "Modeling Word Duration," *Proc. 6th International Conference on Spoken Language Processing(ICSLP)*, Vol.1, pp601-604, 2000.

[8] M.T. Johnson, "Capacity and Complexity of HMM Duration Modeling Techniques," *IEEE Signal Processing Letters*, Vol. 12, No. 5, May 2005.

[9] N. Ma and P. Green, "Context-Dependent Word Duration Modelling for Robust Speech Recognition," *In Proc. Interspeech*, Lisbon, 2609-2612, 2005.