USING RELATIVE DURATION IN LARGE VOCABULARY SPEECH RECOGNITION

M. Jones & P.C. Woodland
Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK.

ABSTRACT

Current large vocabulary continuous speech recognisers (LVCSR) do not model the effects of speech rate on the speech unit durational characteristics. This paper presents work on the investigation of speech rate, presents three durational models which make use of this rate information and the integration of the models into a TIMIT based LVCSR is described. Although TIMIT contains controlled, read speech, with little speech rate variation, experimental work has shown that the relative duration models produce greater word error rate improvements than models which do not take account of the speech rate.

Keywords: Duration modelling, speech recognition.

1. INTRODUCTION

Many large vocabulary continuous speech recognisers (LVCSR) use durational constraints ranging from simple subword unit minimum durations [1] to more sophisticated attempts to model state/unit durations [2]. These constraints can improve recogniser performance by penalising alignments with unlikely durations, arising, for example, from inadequacies in acoustic or lexical modelling. They can also assist in discriminating between recognition units (e.g. *'leek'* has a shorter vowel than the one in *'league'*).

The overall speed at which utterances are produced (the speech rate) obviously effects the durational characteristics of the individual speech units. However, LVCSR systems do not make use of these relative effects. Quantifying rate requires knowledge of the total duration of an utterance section and the number of units produced in that time. This information is unavailable as recognition proceeds.

In earlier work [3] we suggest that features like duration can be exploited using a postprocessing phase after the N-Best algorithm has been used to generate the N most likely utterance hypotheses. The main advantage of the scheme is that utterance level information, such as that needed to calculate speech rates, is made available. This paper discusses the use of speech rate information in phone duration modelling and three possible model schemes are described. Performance changes resulting from the use of the relative duration models are presented. These results are compared with the improvements when using duration models which do not take account of the speech rate.

2. SPEECH RATE

By speech rate, we mean the rate at which individual speech units are produced. This can be expressed as the average unit duration. In deciding how to actually calculate this rate, a number of choices have to made.

The length of the utterance used in the calculation should be long enough to ensure that the rate is representative, while short enough to account for changes in speech rate during the utterance. In our work, we calculate the rate using the entire utterance. This is because the utterances in the TIMIT database are fairly short, with an average of 31 phones.

The speech unit used in the calculation should be such that the variance of the inherent durations of the different units is as small as possible. Otherwise, the rate will be affected more by the content of the utterance than by the actual speech rate. The 'word', then, would be a useless unit to be used in speech rate calculation. In our work we use the phone as the speech rate unit. The syllable was also considered but was rejected due to the greater variance in inherent durations arising from different syllable structures.

Phone durations are normalised before the average phone duration is found. Normalisation causes the speech rate to be affected more by changes in the phones with the smallest variance of absolute durations. This emphases the effect of actual speech rate changes and reduces the impact of other causes of duration variations such as stressing and prosodic phrasing.

Normalised phone duration, τ, is computed as,

$$\tau = \frac{d - \mu}{\sigma} \tag{1}$$

where d is the absolute duration and, μ and σ, the mean and standard deviation of the phone absolute durations found in the training set. After normalisation, the speech rate, r, is found by summing the durations over the utterance and finding the average.

The effectiveness of the average normalised phone duration in differentiating utterance rates was assessed. This was done by computing the weighted average shift of the absolute phone duration distributions for phones found in the slow, average and fast training utterances. The average shift is defined as,

$$\frac{1}{p}\sum_{i=1}^{p}\frac{n_p}{N}\sum_{j=1}^{c-1}\sum_{k=j+1}^{c}\frac{\mu_k^i - \mu_j^i}{\sigma_k^i.\sigma_j^i} \tag{2}$$

where p is the number of phones in the phone set, n_p the number of examples of phone p, N the total number of phones in the training set and c is the number of different utterance classes used in the evaluation (c=3: fast, average and slow utterances). The duration distribution data was gathered in the way described in Section 3.1.

The average shift using average normalised phone duration as the speech rate measure is shown in Table 1.

The shifts achieved using unnormalised phone duration and syllables are also shown for comparison. The larger the shift, the more effective the measure for our purposes. Although in our work we use speech rate in LVCSR, rate

Speech Rate Measure	Avg. Shift x 10^{-3}
Avg. syllable dur.	7.2
Avg. phone dur.	9.8
Avg. normalised phone dur.	12.2

Table 1. Measures of speech rate and the weighted average shift between phone distributions of slow, average and fast utterances.

information has also been used in various other speech processing tasks such as syntactic disambiguation [4] and keyword spotting [5].

3. RELATIVE DURATION MODELS

Three duration models which make use of the speech rate information were built for the hidden Markov model (HMM) phone models used in our speech recogniser. The first modelling scheme uses the utterance rate to partition each phone duration model into a number of submodels while the others use it to adapt a single model. In all cases, the duration training material was phone hand labelled transcriptions from the TIMIT database.

3.1. Partitioned Model

The training utterances were put into three groups, each group containing approximately 700 sentences. The first group contained the fastest utterances and the last the slowest utterances.

For each phone, three submodels were then trained, one for each of the utterance groups. Each model consisted of the absolute minimum/maximum duration observed; the durations between which 90% of durations occurred (i.e. 5% lower/upper limits); a smoothed histogram representation of the distribution of durations using 100 bins of 10ms each; and, the mean/standard deviation of duration for use in a Gaussian model of duration distribution.

To illustrate the differences between models for the different partitions, Table 2 gives some of the statistics for the phone /f/ and Figure 1 shows the envelope of the smoothed histogram duration distributions. The differ-

Rate Group	Min.	Max.	Mean	Std. Dev.
Fastest	13.3	173.7	92.7	28.7
Average	20.2	202.1	101.2	32.1
Slowest	22.6	258.5	114.6	36.3

Table 2. Example statistics from training three submodels for the phone /f/. Absolute duration minima/maxima (Min./Max.); and, duration mean and standard deviations (Mean,Std. Dev.). The three models are for the fastest, average and slowest speed utterances. Durations in *ms*.

ences between the statistics of the rate bands was generally not dramatic. There are two explanations for this. TIMIT is read, controlled speech, leading to less variation in speech rate. Furthermore, no explicit information about the stressing, phrasing, or other duration changing factors, is used in the speech rate measure so no adjustment for these can be made.

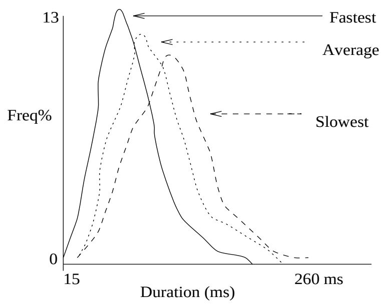

Figure 1. Smoothed histogram of the durations of /f/ found in the fastest, average and slowest utterances.

3.2. Shifted Mean Model

A disadvantage of the partitioned approach is that utterances used to train each partition model have a relatively wide range of utterance rate. Utterances at the top end of one rate band are likely to have durational characteristics similar to utterances at the lower end of the adjacent, succeeding band rather than to those found in the lower end of the same rate band. Increasing the number of rate groups to improve in-class cohesion would lead to a training problem with there being reduced numbers of utterances in each group.

A possible solution to this problem might be to try to adapt an undivided duration model with respect to the speech rate. This adjustment could occur during the scoring of recogniser hypotheses. In this work, we use a very simple adaption scheme. Following [4], the mean of the absolute phone duration, μ, is adapted to give, $\hat{\mu}$, as shown in 3.

$$\hat{\mu} = \mu + \frac{\sigma r}{N} \tag{3}$$

where σ is the standard deviation of the absolute phone duration, r is the sentence rate and N is a scaling factor which was set to give small, cautious shifts without making them negligible as would be the case if N was large.

3.3. Relative Normalised Duration Model

In both of the models described above, the phone durations are absolute ones with the model *forms* being adjusted by the speech rate. In the third model, relative durations are modelled. This approach was considered to attempt to capture directly the effects of the rate on the normalised phone durations while avoiding the rigidity of the partitioned model and the lack of sophistication in the shifting mean approach.

Each normalised phone duration, τ, is transformed into a relative one, τ_r, by,

$$\tau_r = \tau - r \tag{4}$$

Histogram and Gaussian models were generated for the training data.

The distribution of the relative durations was compared with the distribution of durations (τ) which are not relative to the sentence rate. Table 3 gives some example variances of the two statistic types. As was expected, the variances of the relative durations were lower than those where the rate was ignored. The margin was smaller than hoped for, however; the possible reasons for this have been mentioned above.

Phone	Var. τ_r	Var. τ
/ao/	0.84	1.0
/f/	0.90	1.0
/sh/	0.79	1.0

Table 3. Example variances (Var.) for relative (τ_r) durations and those (τ) which do not take speech rate into account. Lower variances in relative case indicate correlation between speech rate measure and phone duration behaviour.

4. INTEGRATING THE MODELS WITH A LVCSR

The duration models are used in a postprocessing phases after the N-Best algorithm has generated a number of hypotheses for each of the test utterances.

4.1. N-Best Algorithm

In speech recognisers using Viterbi decoding the (single) most likely utterance transcription is produced. By contrast N-Best decoding algorithms generate the N most likely hypotheses [6]. The sentence hypotheses are found in order of decreasing likelihood and each has a likelihood score associated with it. In our work, the N-Best front-end recogniser produces word and phone alignments for each transcription.

4.2. Postprocessing

Each of the N transcriptions produced by the front-end recognizer is scored with the duration information. Firstly, a rate is hypothesised for the *transcription*. The way this is then used depends on the model being employed. In the partitioned case, the rate is used to select the appropriate partition of the model for scoring the transcription; for the shifted mean case, the rate adapts the mean of the phone duration distribution; and, with the relative durations model it is subtracted from each normalised phone duration in the transcription to derive a relative value which can be given a log likelihood from the model.

The duration likelihood is then added to the acoustic-phonetic log score. A weight is used to give appropriate emphasis to the duration information, the weight being trained during system development. After the combination of likelihoods, the N-Best list is reordered such that the top transcription has the highest combined score.

5. EXPERIMENTAL WORK

5.1. Baseline System

Experiments were carried out using the TIMIT database. To achieve a large but manageable vocabulary size a subset of the utterances is used which consists of the *sx* type sentences, giving a vocabulary of 1794 words. For the new database two thirds of the speakers in the original TIMIT database were placed in the train set and the remainder made up the test set, with no overlap between sets.

A speaker independent HMM based word recogniser was built using the HTK Toolkit [7]. Monophones were used to build up single pronunciation word models. A word pair grammar acted as the language model constraint. Further details can be found in [3]. Five hundred randomly chosen test sentences were used in the experiments described below. The baseline performance on this set was, 13.62% (word error rate), 29% (sentence error rate). The N-Best algorithm was used to find the 30 most likely transcriptions for every test sentence.

5.2. Applying Relative Duration Information

Each of the three duration models were used to score the transcriptions produced by the front-end recogniser. For the cases where an histogram or Gaussian model was used scoring involved finding the log likelihood of the observed phone duration from the models. However, where boundary statistics were used (i.e. absolute min/max and 5% lower/upper limits) the transcription score was the percentage of phones in the transcription which had durations within the boundaries.

Table 4 details the relative duration models used to score the N-Best lists. The change in recogniser word accuracy, relative to the baseline given in Section 5.1., after N-Best list reordering was recorded. For comparison, the transcriptions were also scored using models which were not adapted for speech rate.

Model Type	Model Params
Partitioned	Abs min/max duration Lower/upper bounds Histogram Gaussian
Shifted mean	Gaussian
Relative durations	Histogram Gaussian

Table 4. Relative duration models used in the scoring of N-Best transcriptions. The transcriptions were also scored using models unadjusted for speech rate.

6. RESULTS

6.1. Effect on Baseline Performance

Tables 5, 6 and 7 show the changes in test set word error rate after scoring/reordering the N-Best transcriptions when using the partitioned, shifted and relative models. The effects of using unadjusted models are also given.

Info Type	Structure	% reduction w.err
Abs min/max	Unadjusted	-0.6%
	Partitioned	2.7%
Lower/upper	Unadjusted	2.2%
	Partitioned	4.1%
Histogram	Unadjusted	6.2%
	Partitioned	10.1%
Gaussian	Unadjusted	5.6%
	Partitioned	9.7%

Table 5. % reductions in word error (w.err) using the partitioned model. 'Unadjusted' results are those obtained with an unpartioned model: no adjustment for speech rate.

Info Type	Structure	% reduction w.err
Gaussian	Unadjusted	5.6%
	Shifted mean	6.4%

Table 6. % reductions in word error (w.err) using the shifted mean and unadjusted models.

6.2. Discussion

As can be seen, the relative duration models outperformed the absolute duration ones in all cases. The better performances when using relative durations can be explained

Info Type	Rate	% reduction w.err
Histogram	Rate Constant	6.1%
	Sentence Rate	7.4%
Gaussian	Rate Constant	5.7%
	Sentence Rate	7.1%

Table 7. % reductions word error (w.err) after applying the relative durations model compared with case when speech rate not used (rate constant).

by considering the scoring of transcriptions in the N-Best list.

Firstly, it is important to remember that a speech rate is calculated for every transcription in the N-Best list. An incorrect transcription may have a similar rate to the correct transcription. However the rate may differ to the that of correct transcription due to erroneous recognition leading to incorrect phone recognition and duration assignments.

Table 8 shows the actual speech rate of two of the test set utterances along with the rates calculated for the first transcription, which is incorrect, and the correct transcription, which occurs further down the N-Best list. After scoring/reordering the N-Best lists using a relative duration model, the correct transcriptions were moved to the top of the N-Best list. In cases such as the first example, where an incorrect transcription and the correct one have a similar rate, the relative duration models have the advantage over non-relative models of being more specific. Where an incorrect transcription has a much different rate from that of the correct transcription, as in the second example, phone durations in the incorrect transcription are scored using data applicable to that speech rate and not the one which the utterance was produced at. As well as penalising incorrectly recognised phones, correct phones should score less well than if they were scored using distributions for the rate they were actually produced at. In this way, durational errors can be amplified by affecting the likelihoods of correct as well as incorrect phones.

Sent. Id.	U.Rate	Inc.Rate	Cor.Rate
mdhl0-5-sx359	-0.22 [1]	-0.19 [1]	-0.19 [1]
fcal1-5-sx357	0.00 [2]	-0.17 [1]	0.00 [2]

Table 8. Test sententces and rates of transcription in hand labbeled data (U.Rate), incorrect and correct transcriptions (Inc.Rate, Cor.Rate). Figures in brackets give submodel relevant to rate, 1 being for the fastest utterances.

The different types of durational information used to score transcriptions bring about varying degrees of improvement in word error rate. The smallest improvements come from using the simple boundary constraints of phone absolute min./max. duration and the 5% lower/upper duration boundaries. This is because of the small number of transcriptions with very unlikely phone duration assignments. The best improvements are achieved using full distribution information (histogram/Gaussian).

Histograms have the potential to model the actual duration distributions more closely than a parametric model. We did achieve better results using histograms than when Gaussian models were employed, but the differences were not significant. Parametric models would be needed if duration models were built for context-dependent subword unit models due to a smaller amount of training data for each duration model.

Of the three relative duration models, the partitioned scheme is the one which produces the best result (10.1% reduction in word error). It accounts for changes in the duration distribution with respect to the speech rate to a much greater extent than the shifted mean model where only the mean is affected. Further work on the shifted model could look at adjusting the variances and finding the optimal (rather than arbitrary) weight given to the rates in shifting. The partitioned model also seems to make better use of the speech rate effects in transcription scoring through using hard partitioning decisions and modelling of absolute durations within the partitions. The other model smooths out these useful contrasts. The relative normalised duration and the shifted models do though have the advantages of being simpler, leading to smaller storage space and computation costs.

7. CONCLUSION

The rate at which utterance sections are spoken affects the duration of the units they contain. Ways of modelling and using these effects in a LVCSR have been investigated with the N-Best algorithm and a postprocessing phase.

Results using a database of controlled, read speech show that greater performance improvements are possible when using the relative models than when using duration models which do not take account of the speech rate. It is believed that these improvements might be more dramatic where there is greater variation in speech rate such as in spontaneous speech.

8. ACKNOWLEDGEMENTS

M. Jones holds a SERC Research Studentship. The authors thank C. J. Leggetter who implemented the N-Best algorithm.

REFERENCES

[1] Paul, D. B. (1991). New results with the Lincoln tied mixture HMM CSR-System. Proc. DARPA Speech and Natural Language Workshop, 65-70, Pacific Grove.

[2] Levinson, S. E. (1986). Continuously variable duration Hidden Markov models for automatic speech recognition. *Computer Speech and Lang.,* Vol. 1. No, 1. 29-45.

[3] Jones, M. & Woodland, P. C. (1993). Exploiting variable width features in large vocabulary speech recognition. *Proc. ICASSP 93*, Vol. 2. 323-326, Minnesota.

[4] Price. P. J., Wightman, C. W., Ostendorf, M. and Bear, J. (1990). The use of relative duration in syntactic disambiguation. *Proc. ICSLP'90*, 1.4.1-1.4.4, Kobe.

[5] Kawabata, T., Hanazawa, T. and Shikano, K. (1988). Word spotting method based on phoneme recognition. ATR Technical Report, TR-I-0065.

[6] Schwartz, R. & Austin, S. (1991). A comparison of several approximate algorithms for finding multiple (N-BEST) sentence hypotheses. *Proc. ICASSP'91*, 701-704, Toronto.

[7] Young, S. J. (1992). HTK V1.4 User, Reference & Programmer Manuals. Cambridge University Engineering Department, Speech Group.