SUPRASEGMENTAL DURATION MODELLING WITH ELASTIC CONSTRAINTS IN AUTOMATIC SPEECH RECOGNITION

Laurence Molloy *Stephen Isard*

Centre for Speech Technology Research,
University of Edinburgh,80, South Bridge, Edinburgh EH1 1HN, GB
http://www.cstr.ed.ac.uk email: lauriem@cstr.ed.ac.uk

ABSTRACT

In this paper a method of integrating a model of suprasegmental duration with a HMM-based recogniser at the post-processing level is presented. The N-Best utterance output is rescored using a suitable linear combination of acoustic log-likelihood (provided by a set of tied-state triphone HMMs) and duration log-likelihood (provided by a set of durational models). The durational model used in the post-processing imposes syllable-level elastic constraints on the durational behaviour of speech segments.

Results are presented for word accuracy on the Resource Management database after rescoring, using two different syllable-like constraint units, a fixed-size N-phone window and simple (no constraint) phone duration probability scoring.

1. INTRODUCTION

Although traditional Hidden Markov Models (HMMs) have proven to be highly successful at acoustic classification they inherit an implausible durational model through the mathematical behaviour of their state transition probabilities. Hidden Semi-Markov Models [7] have partly overcome this problem by replacing the discrete probability associated with a state's self-transition with a continuous duration probability distribution. However, the utility of this treatment of duration is constrained by its assumption of the Markovian principle of independence at the suprasegmental level. This assumption seems to be at odds with previous theoretical studies on segmental duration ([6], [11], [10]) which focus on suprasegmental effects.

More recently, research in the field of speech recognition has focussed on modelling durational behaviour at the suprasegmental level to account for individual theoretical phenomena (e.g. post-vocalic context [4] and speech rate [5]). A more complete account of the full spectrum of theoretically accepted suprasegmental durational effects has also been presented in [9], with a view to improving HMM-based recognition through re-scoring of N-Best sentence output.

In this paper a method of integrating a model of suprasegmental duration with a HMM-based recogniser at the post-processing level is presented. The N-Best utterance output is rescored using a suitable linear combination of acoustic log-likelihood (provided by a set of tied-state triphone HMMs) and duration log-likelihood (provided by a set of durational models). The database used in this task (Resource Management) and the baseline HMM architecture used to obtain the N-Best Lists and acoustic log-likelihood scores are described in section 2.

Post-processing of HMM N-Best lists is not a new technique. However, the durational model used in the post-processing is different in that it imposes syllable-level elastic constraints on the durational behaviour of speech segments. The elastic constraint durational model proposed in this paper is based upon previous work in the field of speech synthesis [3]. The model definition and how it is integrated with the HMM at the post-processing level are discussed in detail in sections 3 and 4.

The proposed model of duration is tested, using a variety of constraint units including two different syllable-like units, a fixed-size window of 3 phones and individual phone units. All the above treatments are repeated for phone sets derived using a number of levels of prosodic subcategorisation. The prosodic contexts for which the phoneme data is subcategorised are chosen according to a stepwise CART tree correlation analysis. The experimental design and the choice of prosodic subcategorisations are discussed in more detail in section 5.

2. MATERIALS

The utility of the proposed elastic constraint model is tested on the Resource Management (RM) database. The database is first separated into 3 speaker-independent sets of utterances for training (3990 utterances), cross-validation (1110 utterances) and testing (900 utterances). A set of multiple mixture, cross-word, (tree-based) clustered state, tied triphones with back-off biphones and monophones is then trained on the training set.

The phonetic segmentation required for the durational model is obtained from the training set using the fully trained HMMs in forced alignment mode and the dictionary used for training and recognition is a *most likely pronunciation* dictionary with a single, fixed phonetic transcription for each word.

In recognition mode, the HMMs are required to produce a list of the 30 best-scoring hypotheses for each test and cross-validation utterance, using a word-pair grammar. The best performance achieved on the test set of utterances is 94.19% word accuracy on the top-scoring hypothesis and 98.73% word accuracy on the most accurate hypothesis out of the top 30.

The latter figure quoted is the maximum accuracy achievable, given only the top 30 hypotheses to choose from. Thus, in relying on an N-Best list as our input for rescoring, we have sacrificed 1.27% word accuracy. In assessing the utility of the duration model proposed in this paper, 98.73% is considered to be a *perfect* score.

3. THE ELASTIC CONSTRAINT MODEL

The concept of elastic constraints is based upon previous work in the field of speech synthesis. The elasticity hypothesis [3] suggests that, within a syllable, phonemes behave like springs of different lengths (mean durations) and elasticities (standard deviations). That is to say that if the syllable lengthens it is hypothesised that all its constituent phonemes should lengthen in proportion to their elasticities. i.e. the ratio of a phoneme's lengthening to its elasticity remains constant throughout the syllable.

In [3], Campbell and Isard state that "*All segments in a given syllable fall at the same place in their respective [duration] distributions*" i.e For any given syllable, there exists a number, K, such that, for all segments within that syllable

$$Duration_{seg} = \mu_{seg} + K\sigma_{seg} \tag{1}$$

For a given speech segment, its **Z-score** is defined as its normalised distance, in units of standard deviation, from its mean duration.

$$Z_{seg} = \frac{dur_{seg} - \mu_{type}}{\sigma_{type}} \tag{2}$$

For a given syllable, its **K-Score** is then defined as the average Z score of the phonemes within the syllable.

$$K_{syll} = \frac{dur_{syll} - \sum_{i=1}^{N} \mu_i}{\sum_{i=1}^{N} \sigma_i} \tag{3}$$

Strict adherence to Campbell and Isard's elasticity hypothesis would imply that all phones within a syllable have the same Z-score, and hence the syllable K-score is exactly equal to all phone Z-scores within the syllable.

A measure of a syllable's deviation from this hypothesis is known as **K-Deviation**. This is defined as the root mean square Z-score deviation from the syllable K-score, calculated over all phones within the syllable. A zero K-Deviation thus corresponds to a syllable's perfect fit to the Elasticity Hypothesis.

$$Kdev_{syll} = \sqrt{\frac{\sum_{i=1}^{N}(K_{syll} - Z_i)^2}{N}} \tag{4}$$

In reality, syllables don't exactly fit the elasticity hypothesis. However, this paper tests the hypothesis that strict elasticity is simply an underlying norm, from which any variance is determined systematically as a result of prosodic contexts. Identifying the main causes of such systematic variance and factoring these effects out of the data may lead to a useful model of duration, based on underlying elastic constraints.

3.1. Methodology

In order to assign a likelihood score to an utterance's durational pattern, a probability has to be assigned to every syllable's K-deviation. It was found in [8] that, for a small database of phonetically rich sentences, syllable K-Deviation was well modelled by a set of gamma distributions, defined solely by the number of phones contained in the syllable. Thus, K-Deviations are calculated for all syllables in the training set, from which a set of gamma distributions are derived.

However, monophonic syllables always exhibit zero K-Deviation. Therefore, an alternative method of scoring is required for these syllables in order that hypotheses aren't favoured purely because of the number of monophonic syllables they contain. To this end, gamma distributions for individual phonemes which occur in monophonic syllables are computed from the training data and all phonemes that constitute monophonic syllables within a given hypothesis are scored against these gamma distributions.

4. POST-PROCESSING

To rescore an utterance, its acoustic log-probability is first linearly combined with the total N-phone syllable K-Deviation log-probability for the utterance.

$$ln(P_{A'}) = (1-\alpha)ln(P_A) + \alpha \sum_{utt} ln(P_N) \tag{5}$$

where

- P_A = probability of acoustic event for an utterance
- P_N = probability of a durational K-Deviation for an N-phone syllable

The resulting log-probability is then linearly combined with the monophonic syllable durational log-probabilities to arrive at a final combined acoustic and durational score.

$$ln(P_{total}) = (1-\beta)ln(P_{A'}) + \beta \sum_{utt} ln(P_1) \tag{6}$$

where

- P_1 = probability of a given duration for a phoneme occurring as a monophonic syllable

In the above calculations, the optimal linear sum weights α and β are defined to be those weights which maximise the word accuracy score on a separate cross-validation set of 1110 utterances.

5. EXPERIMENTAL DESIGN

5.1. Testing the Constraints

In order to test the Elastic Constraint model of duration, the performance of a number of alternative versions was measured

A baseline durational model is constructed, using only phone duration distributions in isolation. The word accuracy of this model allows us to see whether the elastic constraint model has actually achieved anything.

Furthermore, it is possible that elastic constraints exist, but that the syllable as a unit of constraint doesn't buy us any more performance than a non-linguistically motivated fixed-size window of phones would. To test for this possibility, we have tried replacing the syllable in the model with a sliding window of 3 phones.

In both of the above alternative duration models, even though we are moving away from a syllable-like unit of constraint, the prosodic contexts of the phonemes are still largely dependent on our choice of syllable-like unit. We chose to perform all the rescoring in these cases using the maximal onset definition (as defined in section 5.2).

5.2. Syllabification

We have chosen two syllable-like units for inclusion in our model largely because their operational definitions are simple to implement.

Maximal Onset: Each syllable contains a single vowel and the syllable onset length is maximised subject to a set of phonatactic constraints.

Vowel Initial: Each unit begins with a vowel and includes all consonants within the same word up to the next vowel. Word initial consonants are grouped with the unit containing the first vowel in the word.

5.3. Context Dependency

Initially rescoring is performed using models based on context-independent phone duration distributions. However, factors other than the phone identity itself are known to affect phone duration. In order to normalise for these factors, we split the phoneme data pools according to a number of prosodic contexts.

To determine which prosodic factors are important, and in which order they should be applied in this task, for each definition of the syllable-level unit we trained a CART tree [2] on the training set of 146700 phones. Table 1 is a stepwise correlation analysis of prosodic factors which determine phone duration in the training set, using a greedy algorithm, as produced by the resulting CART tree.

Rescoring is repeated using context-dependent phone duration distributions by incrementally splitting the data according to the above factors, in the order specified in the the table 1. In this paper, we only consider the first 4 factors.

6. DEALING WITH SPARSE DATA

One of the major limiting factors inherent in such an exhaustive classification of phonemes is that of data sparsity. In some cases, the gaps in the data may be accidental rather than systematic. It is these accidental gaps which need to be filled in order to score the test data using the durational model.

		cumulative correlation (%)	
Factor	Classes	Onset	Vowel
phonemic identity	39	51.75	51.75
syll position in utt	4	60.06	60.73
this syll stress	2	64.63	64.48
syll position in word	4	67.34	67.18
subsyllabic position	3	69.67	70.91*
next syll stress	3	70.98	69.15*
clustered C?	2	71.98	71.99
prev syll stress	3	72.65	72.64
ambisyllabic C?	2	72.74	N/A**
in stressed word?	2	72.76	72.68

Table 1: Cumulative correlation (using a greedy algorithm) for prosodic factors using 2 different syllable-like unit definitions: (maximal) **Onset** and **Vowel** (initial) units. **next syll stress* and *subsyllabic position* have swapped relative importance for vowel-initial units. ***ambisyllabic C?* is defined for maximal onset units only.

A method of *synthesising* the durational parameters for such gaps has been devised which borrows from Dennis Klatt's work on duration. In the MITalk system [1], *actual* duration is calculated using a linear transformation of an *inherent* duration, according to a number of prosodic rules. In this paper, for all possible pairs of prosodic contexts, a linear transformation for the durational mean of the form

$$\mu_{p_i} = a + b\mu_{p_j} \tag{7}$$

where μ_{p_i} is the mean duration of phone p in context i, is estimated by linear regression. This transformation represents the *average* effect that going from context j to context i has on the durational mean of any given phone duration distribution. In applying transformations from seen to unseen contexts, only reliable linear transformations (correlation ≥ 0.5) are considered.

However, the behaviour of the distribution standard deviation under such a transformation is less easy to model reliably. In the majority of cases, linear regression of context pairs produces minimal correlation with the actual data. Thus, for any given phone, an average standard deviation is computed using all examples of that phone present in the database.

7. RESULTS

Table 2 presents word accuracy and error correction for all rescoring models tested. All models have improved the performance of the baseline HMMs, demonstrating that durational information can be used to correct mistakes made by a HMM in the post-processing stage.

It is encouraging that our best result was achieved using the elastic constraint model with phonetic data partitioned according to utterance finality, with 20% of all possible error corrections being made. The fact that adding extra constraints C & D did not

	A	A+B	A+B+C	A+B+C+D
Phoneme	94.58 **8.59%**			94.98 **17.40%**
3 Phone Window	94.94 **16.52%**			94.99 **17.62%**
Vowel-Initial Syllable	94.88 **15.20%**	94.98 **17.40%**	94.97 **17.18%**	95.04 **18.72%**
Max-Onset Syllable	94.77 **12.78%**	95.10 **20.04%**	95.06 **19.16%**	94.93 **16.30%**

Table 2: word accuracy results for RM test set using different constraint units in the durational model and different levels of syllable-level prosodic context for the phone-set: context-free phoneme identity (A), utterance finality (B), lexical stress (C) and word finality (D). The figures in bold are percentage of possible error corrections made.

further improve the results may implicate the simplistic averaging method proposed for modelling unseen standard deviations. The elastic constraint model makes important use of means and standard deviations.

An improved operational syllable definition might also yield better results.

Acknowledgements

Laurence Molloy is funded by EPSRC grant number GR/K73848

8. REFERENCES

1. Jonathan Allen, M. Sharon Hunnicutt, and Dennis Klatt. *From Text to Speech: The MItalk System*, chapter 9, pages 93–99. 1987.
2. L. Breiman, J. Friedman, and R. Olshen. *Classification and Regression Trees*. Wadsworth and Brooks, Pacific Grove, CA, 1984.
3. W.N. Campbell and S.D. Isard. Segment durations in a syllable frame. *Journal of Phonetics*, 19:37–47, 1991.
4. L. Deng, M. Lennig, and P. Mermelstein. Use of vowel duration information in a large vocabulary recogniser. *JASA*, 86(2):540–548, 1989.
5. M. Jones and P.C. Woodland. Using relative duration in large vocabulary speech recognition. pages 311–314, Berlin, 1993. Eurospeech.
6. D.H. Klatt. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *JASA*, 59(5):1208–1221, May 1976.
7. S. E. Levinson. Continuously variable duration hidden markov models for automatic speech recognition. *Computer Speech and Language*, 1:29–45, 1986.
8. Bernard Payne. Syllable-level timing of segments in speech recognition. Master's thesis, University of Edinburgh, 1994.
9. L. Pols, X. Wang, and L. ten Bosch. Modelling of phone duration (using the TIMIT database) and its potential benefit for ASR. *Speech Communication*, 19:161–176, 1996.
10. A.E. Turk and L.S. White. The domain of accentual lengthening in Scottish English. volume 2, pages 795–798. Eurospeech, 1997.
11. C.W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. Price. Segmental durations in the vicinity of prosodic phrase boundaries. *JASA*, 91(3):1707–1717, March 1992.