CONTEXT-DEPENDENT DURATION MODELLING FOR CONTINUOUS SPEECH RECOGNITION

Tan Lee[1,2], *Rolf Carlson*[1], *and Björn Granström*[1]
[1] Center for Speech Technology, Royal Institute of Technology
S-10044 Stockholm, Sweden
{tan,rolf,bjorn}@speech.kth.se
[2] Department of Electronic Engineering
The Chinese University of Hong Kong, Shatin, Hong Kong
tlee1@ee.cuhk.edu.hk

ABSTRACT

This paper presents a pilot study of using context-dependent segmental duration for continuous speech recognition in a domain-specific application. Different modelling strategies are proposed for function words and content words. Stress level, word position in utterance and phone position in word are identified to be the 3 most crucial factors affecting segmental duration in this particular application. In addition, speaking rate normalization is applied to further reduce the duration variabilities. Experimental results show that the normalized duration models can help improving the rank of the correct sentence in the N-best hypotheses.

1 INTRODUCTION

It is widely admitted that duration of speech segments carry useful and sometimes indispensable information for automatic speech understanding. However, the main-stream ASR technology is weak in duration modelling. In a typical HMM based system, state duration is incorrectly modelled by a geometrical distribution or constrained with simple upper and lower bounds. These methods tend to over-simplify the duration variabilities in human speech and therefore don't provide reliable and significant discrimination for speech recognition.

Many different approaches have been proposed to let duration better contribute to speech recognition. One way is to capture the duration characteristics of specific phones or words and use them as an additional knowledge source for disambiguition. In [1], statistical duration modelling was performed at multiple sublexical levels in a nicely structured hierarchical framework. The duration model was then used to produce probabilistic scores which could be combined with the existing acoustic scores to improve recognition accuracy at both phone and word levels. In [2], a detailed analysis of context-dependent durational variability was performed, from recognition points of view, for the TIMIT database. This study reached a discouraging but realistic conclusion that many duration features are not as consistent as expected and only some of them would be of potential importance for speech recognition purpose.

It is not very surprising that a generic database like TIMIT exhibits great complexity in duration modelling. Many factors may affect the duration of a phone or a word. They include phonetic context, linguistic functionality of the word, sentential position, stress level, rate of speaking, speaker characteristics, etc. Their relative significance differ from one application to another, as shown by K. Bartkova [3].

This paper describes a pilot study on using context-dependent segmental duration for continuous speech recognition in a domain-specific application. The speech recognizer being investigated was designed and implemented for a Swedish spoken dialogue system. Duration of different phones and words are analyzed statistically using hand-labelled training data. The resulting duration models are then used to re-score the N-best sentence hypotheses.

2 THE RECOGNITION TASK – WAXHOLM

WAXHOLM is a spoken dialogue project developed at the Department of Speech, Music and Hearing, KTH, Sweden. The demonstration application is a conversational system on boat traffic and tourist information in the Stockholm archipelago. The WAXHOLM speech database was collected using Wizard-of-Oz techniques. It contains about 1,400 training utterances from 56 speakers and 327 test utterances from another 10 speakers. The vocabulary size is around 850.

Phonetic labelling and segmentation was performed in a semi-automatic way. It means that segmental duration information is readily available to be used for statistical analysis. There are 53 different phones being labelled, including 23 vowels and 30 consonants. The total number of phones are 36,808 and 8,207 in the training utterances and test utterances respectively. Duration analysis described in Section 4 will be dealing with the training data only.

A detailed description of the database can be found in [4].

3 THE RECOGNITION SYSTEM

The speech recognition engine in WAXHOLM has different modes of operation, namely the standard continuous density HMM and recurrent time-delayed neural network (RTDNN) [5]. In this work, only the RTDNN mode is investigated. Throughout this study, we adopt a sparsely connected two-layer network which was realized using the NICO Tool Kit [6] [1].

Forward Viterbi search with beam pruning is performed to produce a wordgraph based on:

1. Phone-level acoustic probabilities estimated by the RTDNN;
2. The WAXHOLM pronunication lexicon;
3. A specially designed word class bigram (perplexity = 28);

Backward A* stack-decoding search is then applied to generate multiple hypotheses of word strings to facilitate higher-level language understanding and dialogue management [7].

In our work, durational information will be used to re-score and re-rank the N-best list. A similar approach can be found in [8] in which the N-best sentence hypotheses given by a conventional HMM recognizer were re-evaluated by stochastic segment model (SSM).

3 DURATION ANALYSIS & MODELLING

Among the 850 words in WAXHOLM, there are 80 function words like "jag" ("I"), "till" ("to"), etc. Function words (FW) are usually quite reduced. This results in: 1) unrealized lexical stress; 2) many pronunciation variations. On the contrary, content words (CW) are pronounced much more carefully with clear lexical stress. As shown in Table 1, the duration of phones in a function word is significantly shorter than that in a content word. This difference is especially noticeable for vowels. Therefore, in our work, duration of CW and FW are modelled separately.

		CW	FW
Total No. of Phones		24809	11999
Average Phone Duration		84	72
Average Vowel Duration	long	163	109
	short	107	81
Average Consonant Duration		66	59

Table 1: A comparison of phone duration (in msec.) in content words and function words in the WAXHOLM database

In addition to the word type, there are many other factors which may affect phone duration. The best known ones include phone position in the word, number of syllables in the word, stress level, post-vocalic consonant and word position in the utterance. These many factors usually co-exist in a hierarchical and interactive way and their relative importance are application dependent [3].

Function Words

The total number of word occurrences in WAXHOM training data is around 9,000. About 48% of them are due to the 80 function words. That is, most function words have high frequency of occurrences and word-level duration modelling can be performed. Word-level modelling, whenever possible, is highly desirable because word duration is more reliable than phone duration.

For each function word, its typical alternative pronunciations are identified from the transcribed corpus. Then word tokens of each pronunciation should be further divided into two sub-classes according to the word position in utterance (final/non-final). Each class is modelled using a normal distribution. However, in the WAXHOLM corpus, almost all function words don't occur at utterance-final position and, therefore, word position becomes a "don't care" in this case.

Content Words

Word-level modelling is not applicable to content words because of their low frequency of occurrences. Context-dependent phone-level modelling is performed instead. Based on previous studies and a preliminary examination of the labeled WAXHOLM data, three major contextual factors are identified as being significant for this particular application. They are organized in a hierarchy as in Figure 1. It is observed that the stress level is most crucial to affect phone duration. The word position in utterance, often referred to as "sentence final lengthening", is also quite important.

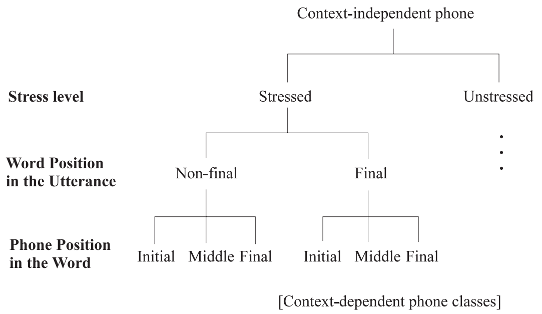
Context-independent phone
Stress level
Stressed
Unstressed
Word Position in the Utterance
Non-final
Final
Phone Position in the Word
Initial
Middle
Final
Initial
Middle
Final
[Context-dependent phone classes]

Figure 1: The three contextual factors for phone duration modelling: stress type, word position and phone position

[1] NICO is a tool kit for Artificial Neural Networks in speech applications. It was developed by Nikko Ström at the Department of Speech, Music and Hearing, KTH, Sweden. For more information and free download, see http://www.speech.kth.se/NICO

Each of the 53 phones is divided into smaller classes with respect to the three factors. Again, each class is modelled by a normal distribution. In case a class doesn't contain enough tokens (e.g. less than ten), it is combined with its nearest neighbour class and the relevant contextual factor becomes a "don't care".

5 SPEAKING RATE NORMALIZATION

For conversational speech in WAXHOLM, the rate of speaking (ROS) varies greatly among different speakers. Even for the same speaker, the ROS may deviate from one utterance to another. In this case, absolute duration doesn't mean too much because of the large unpredictable dynamic range. Duration have to be normalized properly with respect to the ROS.

Let w denote a word in the utterance. If w is a function word, the word-level ROS, denoted by $WROS(w)$, is defined as

$$WROS(w) \triangleq \frac{DUR(w)}{\mu_{DUR}(w)} \tag{1}$$

where $DUR(w)$ is the absolute duration of w in this utterance and $\mu_{DUR}(w)$ is the global average duration (over the whole training database) of w.

If w is a content word and only phone-level duration models are available,

$$WROS(w) \triangleq \text{average}_p(PROS(p)) \tag{2}$$

where p denote the constituent phones of w and

$$PROS(p) \triangleq \frac{DUR(p)}{\mu_{DUR}(p)} \tag{3}$$

Since most utterances in WAXHOLM are quite short (6 words on average), we assume that ROS is stable, to certain extent, within each utterance. Then an utterance-level ROS, denoted as $UROS$ is defined as

$$UROS(w) \triangleq \text{average}_w(WROS(w)) \tag{4}$$

Therefore, $UROS < 1$ for a fast spoken utterance and $UROS > 1$ for a slowly spoken one. Among the 66 speakers, the fastest speaker has an average $UROS$ of 0.722 and the slowest speakers has average $UROS$ of 1.607. This is indeed a wide dynamic range.

For each training utterance, $UROS$ is computed by (1)–(4) and the normalized duration (word or phone) are obtained by dividing the absolute duration by $UROS$. Subsequently, normalized duration models are built in exactly the same way as the absolute duration models.

Table 2 gives a summary of the absolute and normalized duration models for function words. By taking the three contextual factors into account, the standard deviation of average absoluate word duration is reduced by 18% (from 104 msec. to 85 msec.). An additional reduction of 35% (from 85 msec. to 57 msec.) is obtained by speaking rate normalization.

Duration Models					
CI		CD Abs		CD Norm	
$\bar{\mu}_{CI}$	$\bar{\sigma}_{CI}$	$\bar{\mu}_A$	$\bar{\sigma}_A$	$\bar{\mu}_N$	$\bar{\sigma}_N$
225	104	226	85	220	57

Table 2: A summary of duration models for function words in WAXHOLM. μ and σ denote mean and standard deviation (in msec.) of the normal distribution respectively. CI and CD refer to context-independent and context-dependent models respectively

Table 3 gives a summary of the absolute and normalized duration models for context-dependent phones in content words. The standard deviation has been reduced from 40 msec. in the context-independent case to 31 msec. in the context-dependent and normalized case. Greater reduction can be obtained for vowel segments.

	Duration Models					
	CI		CD Abs		CD Norm	
	$\bar{\mu}_{CI}$	$\bar{\sigma}_{CI}$	$\bar{\mu}_A$	$\bar{\sigma}_A$	$\bar{\mu}_N$	$\bar{\sigma}_N$
Long Vowel	138	69	138	61	136	49
Short Vowel	101	42	100	37	100	33
Consonant	64	34	64	31	63	27
All segments	79	40	80	35	79	31

Table 3: A summary of duration models for context-dependent phones in WAXHOLM. μ and σ denote mean and standard deviation (in msec.) of the normal distribution respectively. CI and CD refer to context-independent and context-dependent models respectively

5 N-BEST RE-SCORING

As described in Section 3, the output of speech recognizer in WAXHOLM is in the form of an N-best list, i.e. the N most likely word sequences. Currently the N-best list is generated and ranked with a weighted sum of log acoustic probability and log grammatic probability. Duration information is completely ignored although the phone-level time alignment are readily available for each word sequence.

For each phone and word in a hypothetic word sequence, the absolute duration can be obtained directly from the word graph. Subsequently, $UROS$ can be estimated by (1)–(4) and normalized duration can be computed.

Using the word-level and phone-level duration models, a probabilistic duration score is assigned to each word. If it is a function word, the duration score is obtained, as log probability, from the word-level model with matched sentential position and word pronunciation. If the word is a content word, the duration scores are first computed for all constituent phones from the corresponding context-dependent phone-level models. The word duration score is then equal to the average over all phones.

For the entire word sequence, the overall duration score is computed as the average over all words. Then a combined score is defined as

$$S_{combined} \stackrel{\Delta}{=} w_a \cdot S_a + w_g \cdot S_g + w_d \cdot S_d \qquad (5)$$

where S_a, S_g and S_d denote the acoustic, grammatic and duration scores respectively, and the summation of weights w_a, w_g and w_d is equal to 1.0.

For $N = 30$, the original N-best list is re-ranked using the new combined score. Table 4 shows the effectiveness of the proposed re-scoring method. Different weights have been tried and the average rank of the most desirable sentence is compared. It is observed that using the duration score alone gives better performance than using the acoustic score alone. The reason might be that the segmental duration already reflect, to certain extent, the acoustic realization and grammatical structure of the utterance. It can also be observed that the rank of the most desirable sentence is improved by incorporating the duration score properly.

	w_a	w_g	w_d	Average rank of the most desirable sentence
Case 1	1.0	0.0	0.0	6.32
Case 2	0.0	1.0	0.0	7.13
Case 3	0.0	0.0	1.0	5.89
Case 4*	0.5	0.5	0.0	4.08
Case 5*	0.22	0.19	0.59	3.94

Table 4: A summary of N-best re-scoring results. The "$*$" means that the weights have been optimized using the training data. For case 4, the grammatical score was scaled such that $w_g = w_a$ and the resulted scaling factor is continually adopted for case 5

ACKNOWLEDGEMENT

This work was carried out in and with the support of the Centre for Speech Technology (CTT), KTH, Sweden. The first author would like to thank his colleagues in CTT, namely Nikko Ström, Jesper Högberg, Björn Gämback, Gunnar Fant, Mats Blomberg and Kjell Elenius for their valuable advice.

REFERENCES

[1] G. Chung and S. Seneff, "Hierarchical duration modelling for speech recognition using the ANGIE framework", *Proc. EUROSPEECH-97*, Vol.3, pp.1475 – 1478.

[2] L.C.W. Pols, X. Wang and L.F.M. ten Bosch, "Modelling of phone duration (using the TIMIT database) and its potential benefit for ASR", *Speech Communication*, 19(2), pp.161 – 176.

[3] K. Bartkova, "Some experiments about the use of prosodic parameters in a speech recognition system", *Proc. 1997 ESCA Workshop on Intonation: Theory, Model and Applications*, pp.33 – 36.

[4] J. Bertenstam *et al*, "Spoken dialogue data collection in the Waxholm project", *STL Quarterly Progress and Status Report* 1/1995, pp.28 – 49.

[5] N. Ström, "Continuous speech recognition in the WAXHOLM dialogue system", STL QPSR 4/1996, pp.67-96.

[6] N. Ström, "Phoneme probability estimation with dynamic sparsely connected artificial neural networks", *The Free Speech Journal (http://www.cse.ogi.edu/CSLU/fsj/)*, Issue # 5, 1997,

[7] R. Carlson and S. Hunnicutt, "Generic and domain-specific aspects of the Waxholm NLP and dialogue modules", *Proc. ICSLP-96*, pp.677 – 680.

[8] M. Ostendorf *et al*, "Integration of diverse recognition methodologies through re-evaluation of N-best sentence hypotheses", *Proc. 1991 DARPA Workshop on Speech and Natural Languages*, pp.83 – 87.