NOVEL FILLER ACOUSTIC MODELS FOR CONNECTED DIGIT RECOGNITION

Ilija Zeljković, Shrikanth Narayanan

AT&T Labs, 180 Park Avenue, P.O. Box 971,Florham Park, NJ 07932
{ilija, shri}@research.att.com

ABSTRACT

The context-dependent modeling technique is extended to include non-speech filler segments occurring between speech word units. In addition to the conventional context-dependent word or subword units, the proposed acoustic modeling provides an efficient way of accounting for the effects of the surrounding speech on the inter-word non-speech segments, especially for small vocabulary recognition tasks. It is argued that a robust recognition scheme is obtained by explicitly accounting for context-dependent inter-word filler acoustics in training while ignoring their explicit context dependencies during recognition testing. Results on a connected digit recognition task over the telephone network indicate an improvement in the error rate from 2.5% to 0.9% i.e., about 64% word error-rate reduction, using the improved model set.

1. INTRODUCTION

Modeling of intra-word or intra-phone contexts is widely recognized as an important way to improve recognition accuracy in automatic speech recognition: Context dependent (CD) word and subword models have been shown to outperform context independent (CI) models [1, 2]. Most current acoustic modeling schemes, however, ignore the effects of the surrounding speech on the non-speech regions. However, it was well demonstrated in [3] that inter-word coarticulation has significant effect on recognition performance: A 15-25% reduction in word error rate over a system without inter-word units was reported for the DARPA resource management task. However, for small vocabulary tasks such as connected digit recognition, we argue that modeling the inter-word coarticulation in the training process and ignoring it in the recognition phase contributes significantly to robustness of the recognizer since it is assumed that there is enough discriminative information in the word kernels for recognition. We also argue that these inter-word contexts are heavily influenced by background noise and speaker's non-verbal acoustic productions (breath, lip and tongue smacks, etc) hence discounting the need for *exact* context-specific inter-word modeling. In summary, the influence of speech (and environment/speaker noise) variations on non-speech segments in connected speech results in decreased robustness, indicating the need for good filler models for the background acoustics.

In this paper, we demonstrate that filler HMMs for non-speech regions that incorporate context effects of the surrounding speech segments provide significant improvement in recognition accuracy and increased robustness for small vocabulary tasks. The word error rate for a connected digit recognition task over the telephone network showed a decrease from 2.5% to 0.9% for the case with and without context-dependent filler models. In addition, we also report a simple alternate strategy for modeling filler acoustics that use low-resolution word models themselves as pseudo-fillers. Although these proposed filler models *do not* contribute directly to word discrimination, they greatly eliminate the influence of intra-word speech and non-speech noise from the discrimination process, thereby significantly contributing to the overall accuracy and model robustness to speaker and

background variations.

The paper is organized as follows: Details of acoustic modeling are given in Sec. 2 and the results are reported in Sec. 3.

2. ACOUSTIC MODELING METHODOLOGY

Assuming a head-body-tail modeling scheme [2, 4], it can be noted that the heads and tails of each word unit model the contextual effects. In addition, in our proposed scheme, we model intra-word non-speech segments as head-body-tail units, attempting to capture the contextual influence of the speech segments on the acoustics of the non-speech regions. The belief here is that these filler HMMs help model the effects of the speaker and speech styles present in natural connected speech. Furthermore, under this rationale, the scope of the filler HMMs would be further generalized if, during the recognition process, these filler HMMs were not tied to specific words or contexts but rather used as free fillers that could be embedded in any intra-word non-speech region.

Finally, as a simple alternate strategy to the above mentioned filler models that explicitly model contextual effects, a new filler model set comprising low-resolution word HMMs is also proposed. The goal here is to reduce error in cases where there is confusion between the speech and background acoustics. These low-resolution (pseudo) filler models maybe used to enhance the conventional CI background model, maybe with some loss in performance, if the use of CD fillers is deemed prohibitive (in terms of memory).

In summary, the recognizer using the novel filler HMMs is expected to be more robust to both speaker and environment noise variations, since the intra-word acoustics are much better modeled.

2.1. MODELS: DIGITS AND FILLERS

Speech units (words and subwords) as well as background acoustics in our HMM-based recognizer are modeled by first order, left-to-right HMMs with continuous observation densities. The observation vector consists of 39 features: dynamically normalized energy, 12 LPC derived cepstral coefficients, as well as their first and second derivatives.

2.1.1. CONTEXT INDEPENDENT MODELS

Digit Models: Eleven digits, including '*oh*' and '*zero*', were used in the evaluation task. Each digit was modeled with a 20 or 15 state HMM and 16 Gaussian mixtures.

Background Models:
The conventional CI background model was created from the silence segments using a single state HMM with 128 Gaussian mixtures.

2.1.2. PSEUDO FILLER MODELS

One way of increasing the scope of the conventional CI silence/background model, is to introduce new low-resolution word models that serve the purpose of fillers that are built using the same speech data used to obtain the digit models. The goal of these low-resolution filler models is to take over cases prone to insertion errors due to acoustic confusion between the foreground speech and the background acoustics. These models were constructed as single state HMMs with 2 Gaussian mixtures and are used as filler models in a word spotting mode during testing i.e., these models are allowed to loop in the beginning and the end of the digit strings. In the next section, we propose filler models explicitly derived from background acoustics.

2.1.3. CONTEXT DEPENDENT MODELS

A straightforward way to model context dependent digit variations is to build a separate digit model for each possible combination of left and right contexts. For connected digit recognition task, this implies that every digit would have 144 models, which may be prohibitive in terms of memory and speed requirements for realtime applications.

Digit Models:
Instead of making a full model for each digit in each context, one can consider that a digit d^i is made up of a head h^i, body b^i, and tail t^i and that only digit's head and its tail are affected by the preceding and following contexts, respectively. This implies that the digit body can be shared for all contexts of the same digit. In summary, each digit d^i is represented as:

$$d^i = h^i b^i t^i \quad (1)$$

If digit d^i is preceded by digit d^j then only its head will be affected by d^j and will be represented by model h^i_j. Similarly, the following digit d^k will affect only the tail of digit d^i. The tail of digit d^i is specific for the following digit d^k and is denoted by t^i_k. The context dependent model for digit d^i with preceding context d^j and following context d^k is given by:

$$d^i_{jk} = h^i_j b^i t^i_k \quad (2)$$

Specific heads and tails for each digit are made for the silence context as well. In this case, the total number of HMM units for each digit is equal to 25: 12 heads (for each of 11 preceding digits and 1 silence), 1 body, and 12 tails (one for each possible following context). In this paper, heads and tails are modeled with 3 state HMMs and digit bodies have 14 or 9 HMM states. Probability density distribution for each state is modeled with 16 Gaussian mixtures.

Background models: Traditionally, silence has been modeled with a context independent HMM even when digit HMMs were context dependent [1]. In this study, context independent single state *silence* HMM is compared with context-dependent silence HMM (a.k.a background or inter-word contexts) having the same head-body-tail form as digit HMMs

$$s = h^s b^s t^s \quad (3)$$

Head and tail are specific for each preceding context (digit or silence) and following context, respectively. For example, silence s_{jk} between digits d^j and d^k is modeled by

$$s_{jk} = h^s_j b^s t^s_k \quad (4)$$

Each silence head, body and tail are modeled by a single state HMM with 16 Gaussian mixtures.

Explicit silence context dependency with respect to specific digits is utilized only during model building at the training stage. We found that the recognition performance is more effective if the filler models are treated without being constrained to specific digit contexts. This is reasonable because the goal of these new filler HMMs is to capture general contextual effects of speech on non-speech regions prevalent in connected speech. Hence, the decoding network during the testing allowed each of 25 silence single state HMMs to be freely embedded before/after any other unit. Such a procedure, as will be shown in the next section, yields a fairly robust recognition scheme.

3. RECOGNITION EXPERIMENTS

In this section we describe the details of the model training followed by discussion of the test results.

3.1. TRAINING DATA

The HMMs were trained on the data extracted from four different databases collected over the telephone network. Two of those databases were collected during special speech data collection efforts while two other databases were collected during actual AT&T service field trials. Various HMM model sets were created by using combinations of context independent and context dependent digit model and background models in addition to pseudo filler models.

3.2. TEST RESULTS

The algorithm was tested on speech data collected from field trials of three telephone-based speech recognition services. Total number of digit strings used in testing is 11000. Number of digits in each utterance varied from 4 to 16. No constraints were posed on the length of the digit strings during the recognition process (i.e., unknown length grammar). Any digit sequence, including zero-length,

could be decoded by the recognizer.

Table 1 summarizes the recognition results for various combinations of digit and silence context dependency. The conventional (baseline) case is when silence is context independent while the digit models may be context independent or context dependent. The first method of improved filler modeling is to use low-resolution pseudo-filler models (Sec. 2.1.2) which provides anywhere between 35% to 50% reduction in error rates depending on whether CI or CD digit models are used. It is, however, interesting to note that there is higher number of cases where there were no solution found or the entire string was deleted (under a given search width, which was fixed for all the experiments reported) when pseudo-fillers are used in conjunction with CD digit models indicating the inherent limitations of this type of filler modeling that results in classification difficulties during decoding.

The second method is to use context-dependent silence models Sec. 2.1.3. When context dependency for silence is denoted by 'yes', it implies that context dependent silence models are trained, but each unit (head, body or tail) is used in a context-independent fashion during testing and allowed to be freely embedded at any node in the decoding network. The results clearly show the importance of context dependent filler models. The best performance is achieved when both digits and silence are rendered context dependent. However, the performance suffers much less if CD silence is used together with CI digits rather than, CD digits with CI silence (Table 1). This suggests that contextual variations in digit models are less crucial for digit-discrimination when compared to the need for capturing speech effects on non-speech regions for providing increased robustness.

HMM's Context Dependency:		Additional Pseudo	**Performance (%)**	
Silence	Digit	fillers	Error	NS
NO	NO	NO	3.9	5.6
NO	NO	YES	2.52	2.51
NO	YES	NO	2.74	2.23
NO	YES	YES	1.31	4.6
YES	NO	–	**1.12**	**.02**
YES	YES	–	**0.90**	**.03**

Table 1: Word error rates for connected digit recognition (telephone channel data) with and without 'context dependent' filler models. Also included are results when additional pseudo-filler models are used to supplement CI silence models. 'NS' denotes cases where no solution was found for a given beam width or the entire string was deleted, maintained the same for all the experiments.

4. CONCLUSIONS

Context dependent acoustic models for digits provide improved performance over context independent models. In this paper, we show that context-dependent filler modeling further improves recognition performance significantly (about 64% reduction in error rate). The proposed method of modeling intra-word acoustics during training (so as to eliminate their adverse effects on the model set that does not explicitly account for them) but ignoring their explicit context dependencies during the recognition decoding was found to yield increased robustness for small vocabulary tasks. Furthermore, an alternate strategy for constructing filler models using low resolution word models is proposed for cases where the use of a full context dependent filler set may be prohibitive. The effectiveness of the proposed strategies to speaker and background variations are well demonstrated by results on connected digit recognition tasks.

5. REFERENCES

[1] Wilpon J.G. et al, "Connected Digit Recognition", *BL Technical Memorandum* , November 1994

[2] Lee, C.-H., Giachin, E., Rabiner, L. R., Pieraccini, R., and Rosenberg, A. E., "Improved acoustic modeling for large vocabulary continuous speech recognition," *Computer, Speech & Language*, Vol. 6(2), pp. 103-127, 1992.

[3] Giachin, E., Lee, C.-H. Rabiner, L. R., Rosenberg, A. E., and Pieraccini, R., "On the use of inter-word context-dependent units for word juncture modeling," *Computer, Speech & Language*, Vol. 6(2), pp. 197-213, 1992.

[4] Zeljkovic I. and Narayanan S., "Improved HMM phone and triphone models for realtime ASR applications", Proc. of ICSLP, pp. 1105–1108, Oct 1996.