SUB-PHONETIC MODELING FOR CAPTURING PRONUNCIATION VARIATIONS FOR CONVERSATIONAL SPEECH SYNTHESIS

Kishore Prahallad, Alan W Black and Ravishankhar Mosur†

{skishore, awb, rkm+}@cs.cmu.edu
Language Technologies Institute, Carnegie Mellon University, USA
†Institute for Software Research International, Carnegie Mellon University, USA

ABSTRACT

In this paper we address the issue of pronunciation modeling for conversational speech synthesis. We experiment with two different HMM topologies (fully connected state model and forward connected state model) for sub-phonetic modeling to capture the deletion and insertion of sub-phonetic states during speech production process. We show that the experimented HMM topologies have higher log likelihood than the traditional 5-state sequential model. We also study the first and second mentions of content words and their influence on the pronunciation variation. Finally we report phone recognition experiments using the modified HMM topologies.

1. INTRODUCTION

Modeling of pronunciation variations in conversational speech is essential for speech recognition as well as speech synthesis. The state-of-art speech synthesis systems are built using unit selection databases of carefully read speech recorded in a controlled environment. While these systems produce high quality natural speech they produce little effect of a conversation and lack the genre and style of conversational speech.

To build a speech synthesis system imitating conversational style, it is necessary to model the pronunciation variations that occur in conversational speech. It is typically noticed that the conversational speech is shorter in duration and less accurately pronounced than the careful read speech or over articulated speech as in the case of drama or story telling scenarios. It is also observed that highly probable words in a given context are spoken less accurately than the less probable words. A similar effect could be observed in the case of first mention (occurrence) of a content word and its subsequent mentions in the rest of the conversation.

Less accurate pronunciation of speech corresponds to shortening of the word durations and usage of alternate pronunciation variations. There has been recent efforts in speech synthesis to model pronunciation variations. Werner et al., used n-gram language model to predict word durations and segmental pronunciation [1]. Miller trained neural network models using syntactic and prosodic information to predict the pronunciation variations [2]. Jande used phonological rule system for adapting the pronunciation for faster speech rate [3]. Bennett et al., used acoustic models trained on single speaker database to label the alternate pronunciations of the words: "to, for, a, the" and used CART tree to predict the probable pronunciation with the given context [4].

There has been considerable research in speech recognition field towards capturing the pronunciation variants. Bates *et al.*, showed that prosodic features derived from energy, F0 and duration could be cues to model the pronunciation variability [5]. Nedel *et al.*, used phone splitting technique to model the pronunciation variants of two phones AA and IY [6].

Most of the work in speech recognition and speech synthesis use multiple entries in the dictionary generated either manually or by automatic means. Typically an alternate entry of a word is generated by deletion, insertion and substitution of the phones in the base form of the word. This type of modeling makes a binary decision implying that the base form of a word undergoes a complete change as described by its pronunciation variant. Recent studies have shown that a phone is not completely deleted or substituted but is modified only partially. Proposed solutions to model this change include state level pronunciation models where base form model shares the Gaussian densities with its alternate pronunciation form [7].

In this paper, we extend our views on the fuzzy morphing of the phones during conversational speech. Traditionally a phone is modeled using three or five state Hidden Markov Model (HMM), with a reason that the speech production system goes through two transient states (first and last) and a stable state (middle). Arcs which skip the middle states are often used to relax the 3-state restriction. This type of models are useful when the definition of a phone and its variants is crisp enough to list all the possible variants and their occurrences. However, the practical issues and the difficulties surrounding the clearer transcription, phone set and the use of 3-state sequential models could be observed in [6] [7] [8] [9] [10]. To model the conversational speech, it is important to address the fundamental issue of modeling a phone. From the speech production point of view, it seems more plausible to

associate a phone with a set of production events which could be treated as sub-phonetic states. The speech production system chooses a subset of these sub-phonetic states to produce a phone. However the choice of these sub-phonetic states and their sequence may depend on the nature, style, context and environment of the speaker. In this paper we experiment with two different topologies to model the insertion and deletion of sub-phonetic states during realization of a phone. We refer to these two different topologies as Mod1 and Mod2 and compare with the traditional 5 state sequential topology which is referred to as Mod0.

This paper is organized as follows: Section 2 describes the models Mod1 and Mod2 and specify the motivation to choose these topologies. Section 3 discusses our experimental results on synthesis database. Section 4 provides the experiment results on phone recognition using Mod0, Mod1 and Mod2.

2. MODELING SKIPPING AND INSERTION OF SUB-PHONETIC EVENTS

The primary motivation for experimenting with different HMM topologies for a phone model is to allow the acoustic models the capabilities and flexibility to imitate the speech production process. The traditional three or five state sequential models has several constraints to imitate reduced and full form of the sounds. In order to have the flexibility of skipping and insertion of sub-phonetic states, we experiment with two different topologies of HMMs as shown in Fig. 1. Mod0 is the traditional 5-state sequential phone model. In Mod1 all the states within a phone are fully connected, i.e., transition from any state to any other state is allowed. Any state could be act as beginning state or an ending state. By allowing all possible connections in Mod1, the phone can choose any number and any sequence of sub-phonetic states to model the speech signal. Thus insertion and deletion of sub-phonetic could be handled easily. In Mod2 a state is connected to all other states within a phone but only in the forward direction. While Mod1 is a completely relaxed model of a phone, Mod2 has lesser restrictions than Mod0.

3. EXPERIMENTAL RESULTS AND CUES FOR PRONUNCIATION VARIANTS

To investigate the usefulness of Mod1 and Mod2 over Mod0, we have used the FAF database [11]. Each utterance in this database is a short story spanning multiple sentences. Unlike Arctic databases where there is some effort by some of the voices to render a consistent pronunciation across the utterances, FAF utterances were recorded by a male speaker in a fluent tone to have natural pronunciation variations. In total, there are 107 utterances, consisting of over 14,000 words. They were recorded by a male native speaker of American English from a Midwest American region. This database had

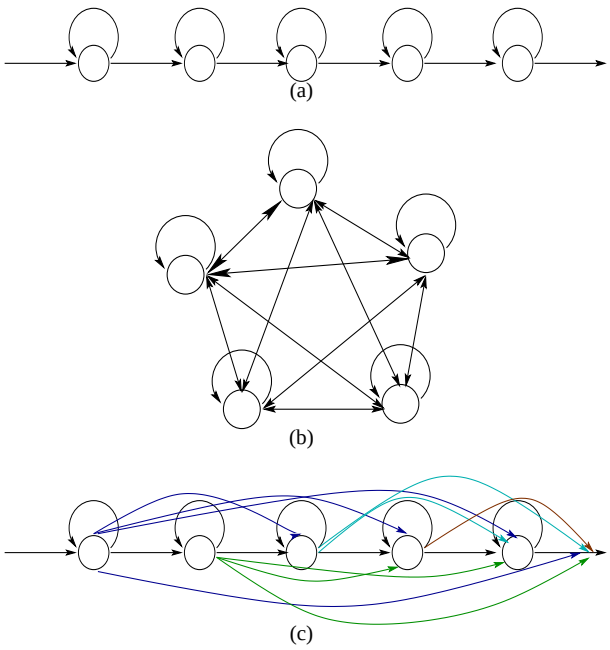

Fig. 1. (a) Mod0: Traditional 5-state sequential phone model (b) Mod1: Fully connected states inside a phone model (c) Mod2: Forward connected states inside a phone model. *Note that in Mod1 and Mod2, any state could be a beginning state or an ending state.*

been designed to capture the supra-sentential prosodic variation and to capture the prominence associated with the topic words. The utterances had a mean length of 45.4 seconds, and a standard deviation of 16.6 seconds. Thus the database is roughly about 80 minutes of speech.

We performed speech segmentation of FAF database using Mod0, Mod1 and Mod2. It should be noted that all three models Mod0-2, consists of 5 states per phone and 1 Gaussian per state. Thus they have same number of parameters. The only difference lies in the fashion in which the states are connected to each other inside a phone model. Feature extraction was performed on FAF utterances to derive Mel-Frequency Cepstral Coefficients (MFCC) for every 5 ms. The HMMs models were trained on FAF database using forward-backward algorithm. To perform the segmentation, the utterances were force aligned to the sequence of phone HMMs and Viterbi decoding algorithm was used to mark the phone level and word level boundaries. The average log likelihood scores obtained for these utterances using Mod0, Mod1 and Mod2 are show in Table 1. It is clear that the average log likelihood scores of Mod1 and Mod2 are better than Mod0 thus indicating a better fit for the speech data.

Table 1: Average log likelihood scores of utterances of FAF database using Mod0, Mod1 and Mod2

Model	Avg. Log Likelihood
Mod0	-24217
Mod1	-23522
Mod2	-23978

Table 2. shows further evidence by displaying the log likelihoods of the three different words for these three models Mod0-2. From Table 2., it can be observed Mod1 and Mod2 seem to model the data in terms of likelihood scores better than Mod0.

Table 2: Log likelihood scores of words w.r.t Mod0, Mod1 and Mod2. The number inside the parenthesis of the word indicates its mention.

Word	P(W/Mod0)	P(W/Mod1)	P(W/Mod2
Ireland (I)	-2836	-2771	-2665
Ireland (II)	-2987	-2650	-2528
France (I)	-2112	-2036	-1254
France (II)	-2247	-2180	-2317
Australia (I)	-4617	-4017	-3849
Australia (II)	-3242	-2708	-2969

As FAF database was designed and recorded keeping in mind the prominence associated with the topic/content words, we specifically looked at the state sequence obtained for the first and second occurrences (mentions) of the words in an utterance. Our assumption is that first mention of a word in a given conversation is relatively better articulated than their second mentions. Thus we were hoping to see different state sequences across first and second mentions, but consistent state sequences within the first/second mentions. Table 3. shows the state sequence obtained for the first and second mentions of the word *Ireland*. This example was picked manually and was one of the consistent examples matching our intuition.

Table 3: State Sequence obtained for the I and II mentions of the Word *Ireland*

Utt. Id	Mention	State Sequence
197	I	242 243 241 101 247 248 247 8 211 210 207 208 23
196	I	242 243 241 101 247 248 247 8 211 210 207 208 23
197	II	242 243 242 243 241 101 247 248 247 8 12 210 208 23
196	II	242 243 241 243 100 247 248 8 211 207 208 23

Further evidence of pronunciation variance with first and second mention of content words was found by building duration models from the same FAF database. In addition to our standard features of phoneme type and context, position in syllable, word and phrase etc. We introduced an additional feature `MENTION` that marks content words, stating the number of mentions within the current paragraph so far. All function words (non-content words) were marked with 0, the first mention of a content word was marked with 1, the second mention of that word was marker with 2, and so on. We used a standard CART tree technique to build duration models. On held out test set of 4560 samples (10%) we achieved the following results.

	RMSE	Correlation
without `MENTION`	0.876	0.458
with `MENTION`	0.869	0.497

The Root Mean Square Error (RMSE) is give in phoneme dependent z-scores. The low RMSE score and relatively higher correlation value in the second column suggest the usefulness of ”with MENTION” for pronunciation variation.

4. PHONE RECOGNITION USING MOD0, MOD1 AND MOD2

In order to compare the modeling capabilities of Mod0, Mod1 and Mod2 for recognition, we performed phone recognition experiments. During recognition the phone models were fully connected to each other thus forming an ergodic sentence HMM. There was no use of phone-phone transition probabilities or any language model in these experiments. The hypothesized phone sequence by the recognizer was compared to the original phone sequence and the comparison is reported in terms of Phone Error Rate (PER). The phone recognition performance obtained for each of these models is shown in Table 4.

Table 4: Performance of Phone Recognition using Mod0, Mod1 and Mod2.

Model	Acc.	PER	Ins.	Del	Sub
Mod0	50.3%	56.9%	3699	8926	16320
Mod1	50.3%	90.5%	20749	4123	21028
Mod2	49.0%	106%	27998	2698	23141

The phone recognition results of Mod0 mentioned in Table 4. should be read with the note that we have used static MFCC features and did not make use of their delta coefficients. At the same time for better comparison purposes we have not used more than one Gaussian per state. The other characteristic of this experiment is the use of single speaker data.

From Table 4. it can be observed that Mod1 and Mod2 perform poorly in comparison with Mod0 inspite of increased likelihood fit to the data. It should be noted that the increased likelihood was obtained during the force alignment. The constraint here was the apriori knowledge of phone sequence and given this constraint it may have been advantageous for Mod1 and Mod2 to be relaxed enough internally to select different sub-phonetic states. However, in the case of recognition we are trying to deduce the phone sequence. This search process needs some constraint which is better found in the sequential model Mod0 as compared to Mod1 and Mod2. Given already relaxed models Mod1 and Mod2, the decoding search process do not seem to have enough constraints thus giving rise to many insertions.

5. CONCLUSION

In this paper we have experimented with fully connected state models and forward connected state models and shown that these models have better log likelihood scores than the traditional 5-state sequential models. The motivation to use a

different topology is to emulate the insertion and deletion of sub-phonetic states as done by the speech production process. We have also reported the effect of first and second mentions of the word for pronunciation variation.

Pronunciation modeling is essential for building conversational speech synthesis system. It is also a well studied field in the speech recognition. However there seems to be some modeling differences in synthesis and recognition. A speech synthesis system is always presented with a phoneme (or Grahame) sequence to synthesize. Given this constraint or prior information relaxed models such as Mod1 and Mod2 may be used for better modeling of the data. However in the case of speech recognition the constraint models such as Mod0 seem to guide the search process better.

6. ACKNOWLEDGMENTS

We would like to thank Dr. James K Baker for the discussions, suggestions and comments on this work.
This work was supported by the US National Science Foundation under grant number 00205731 "ITR: Prosody Generation for Child Oriented Speech Synthesis". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7. REFERENCES

[1] Werner S., Eichner M., Wolff M., and Hoffmann R., "Toward spontaneous speech synthesis - utlilizing language model information in TTS," *IEEE Trans. Speech, Audio Processing*, pp. 436–445, 2004.

[2] Miller C., *Pronunciation Modeling in Speech Synthesis*, PhD dissertation, University of Pennsylvania, 1998.

[3] Jande P.-A., "Phonological reduction in swedish," in *Proceedings of ICPhS*, 2003, pp. 2557–2560.

[4] Bennett C.L. and Black A.W., "Prediction of pronunciation variations for speech synthesis: A data-driven approach," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Philadelphia, Pennsylvania, 2005.

[5] Bates R. and Ostendorf M., "Modeling pronunciation variation in conversational speech using prosody," in *Proceedings of ISCA Tutorial and Research Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language*, Colorado, USA, 2002.

[6] Nedel J.P., Singh R., and Stern R.M., "Automatic subword unit refinement for spontaneous speech recognition via phoneword splitting," in *Proceedings of Int. Conf. Spoken Language Processing*, Beijing, China, 2000.

[7] Saraclar M. and Khudanpur S., "Pronunciation change in conversational speech and its implications for automatic speech recognition," *Computer Speech and Language*, vol. 18, no. 4, pp. 375–395, 2004.

[8] Hain T., *Hidden Model Sequence Models for Automatic Speech Recognition*, PhD dissertation, University of Cambridge, 2001.

[9] Magimai Doss M., *Using auxiliary sources of knowledge for Automatic Speech Recognition*, PhD dissertation, Swiss Federal Institute of Technology Lausanne (EPFL), 2005.

[10] Yu H. and Schultz T., "Enhanced tree clustering with single pronunciation dictionary for conversational speech recognition," Geneva, Switerzland, 2003.

[11] Zhang J., Toth A., Collins-Thompson K., and Black A., "Prominence prediction for super-sentential modeling based on a new database," in *Proceedings of 5th ISCA Speech Synthesis Workshop*, Pittsburgh, USA, 2004, pp. 203–208.