SIGNIFICANCE OF EARLY TAGGED CONTEXTUAL GRAPHEMES IN GRAPHEME BASED SPEECH SYNTHESIS AND RECOGNITION SYSTEMS

Gopala Krishna Anumanchipalli†, *Kishore Prahallad* † ‡ *and Alan W Black* ‡

‡ Language Technologies Institute, Carnegie Mellon University, USA
†International Institute of Information Technology, Hyderabad, India
gopalakrishna@research.iiit.ac.in {skishore,awb}@cs.cmu.edu

ABSTRACT

In this paper we present our argument that context information could be used in early stages i.e., during the definition of mapping of the words into sequence of graphemes. We show that the early tagged contextual graphemes play a significant role in improving the performance of grapheme based speech synthesis and speech recognition systems.

Index Terms— Grapheme, Speech Synthesis, Speech Recognition, Contextual Graphemes, Minority Languages

1. INTRODUCTION

Pronunciation dictionaries define the mapping between the words and basic sounds of a language and thus play a vital role in building speech synthesis and speech recognition systems. Fig. 1 shows the schematic dependence of pronunciation and linguistic knowledge for building speech systems. However, there exist many languages where such linguistic resources aren't available to build speech synthesis and speech recognition systems. For languages which don't have such pronunciation dictionaries, one way to obtain this resource is using the language expert(s) and generate the resource manually. Compilation of such resources in the required format and size takes time as well as requires large capital investment [1]. In some situations, it is even difficult to find an expert in the language area to manually create the required information. Languages for which linguistic resources are scarce or not available are referred to as minority languages.

To build speech synthesis and speech recognition systems in minority languages, techniques starting from basic grapheme based approaches to extraction of linguistic information with the aid of acoustic data have been developed [2] [3] [4]. In grapheme based speech recognition and speech synthesis systems, grapheme is used as basic unit and thus the pronunciation of a word is mapped to sequence of graphemes. For example, words such as "cat" and "church" are mapped to sequence of graphemes "c a t" and "c h u r c h" respectively. Such mapping is used to build Hidden Markov Models (HMM) models for each grapheme by forced-alignment and iterative estimation of the parameters of the models. It is easy to see that the grapheme "c" has more than one pronunciation such as in *cat* and *church* and thus the models built for graphemes are likely to be gross and ambiguous. To resolve the ambiguity, context information is used in the form of previous and next grapheme in *later stages* to build context-dependent models in speech recognition algorithms and to cluster the units using context information in speech synthesis [5].

Such grapheme-based systems have been proposed before for both synthesis [2] and recognition [3]. These techniques have shown promise but even in languages where the relationship between the orthography and the phonetics are fairly transparent, there are still complexities that make those systems not quite as good as a pronunciation based on phones. It is that difference between simplistic grapheme based systems and rich phonetic systems that we wish to reduce.

In this paper we present our argument that context information could be used in early stages i.e., during the definition of mapping of the words into sequence of graphemes. We show that the early tagged contextual graphemes play a significant role in improving the performance of grapheme based speech synthesis and speech recognition systems.

2. MOTIVATION TO USE EARLY TAGGED CONTEXTUAL GRAPHEMES

Typically speech recognition and speech synthesis systems have pronunciation dictionaries to handle standard words and a grapheme-to-phoneme model to handle new words such as proper nouns etc. To model the grapheme-to-phoneme relationship, a grapheme and its 2-level context (previous 2 graphemes and next 2 graphemes) is used to build Classification and Regression Trees (CART) in supervised mode to predict the corresponding phone. Given a grapheme sequence CART exploits the context information present in the grapheme sequence and predicts the corresponding sequence of phones which are then aligned with the acoustic data to build phone level models.

It could be observed that one could remove the CART

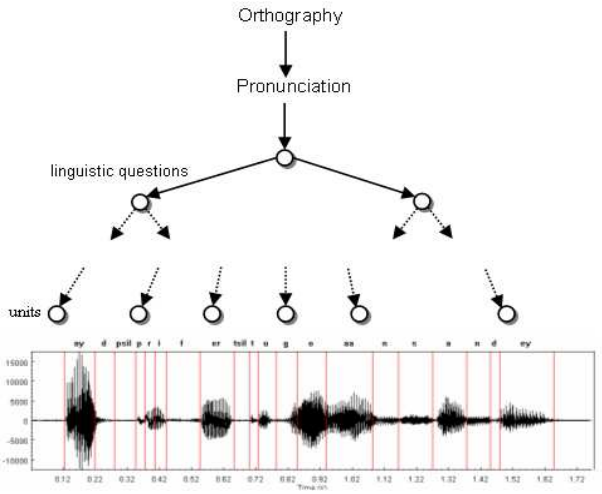

Fig. 1. Dependence on pronunciation and linguistic knowledge for speech processing.

model from the process described above and allow context tagged graphemes to align with the acoustic data. This leads to tagging of the graphemes with the context information in the early stages *as opposed to* exploiting the context information in later stages as done in typical grapheme based speech recognition and speech synthesis systems.

Early tagging of context information to the graphemes would effectively define the mapping of words "cat" and "church" as "#ca cat at#" and "#ch chu hur urc uch ch#". Here # denotes beginning or ending of the word. Note that the early tagging of contextual graphemes is different from context-dependent modeling of graphemes in later stages of speech recognition and speech synthesis. The later stages of context dependent modeling could still be applied to these early tagged graphemes.

3. DATABASE AND EXPERIMENTAL SETUP

For all experiments reported in this paper, we have used RMS voice from ARCTIC database [6]. The database consists of 1128 utterances spoken by a US English male speaker. To validate the early tagging of context information to the grapheme, we have restricted ourselves to 1-level context, i.e., use of the immediate left and immediate right grapheme to tag the current grapheme. Thus we refer to this 1-level context grapheme as trigrapheme in this work. A grapheme with 0-level context is referred to as unigrapheme. In order to validate the potential of trigrapheme units, we have conducted experiments in both speech synthesis and speech recognition and compare the performance of trigraphemes, unigraphemes and phones. In speech synthesis we have built unit selection voices and conducted perceptual study to evaluate the performance of trigrapheme, unigrapheme and phone units. In speech recognition, we have built a phone decoder, grapheme decoder and trigrapheme decoder on RMS voice and measured error rate in terms of deletions, insertions similar to word error rate used in speech recognition systems.

4. TRIGRAPHEME BASED SPEECH SYNTHESIS SYSTEM

For experiments in synthesis, unit selection voices are built using the FESTVOX framework [7]. Two separate voices corresponding to unigrapheme and trigrapheme based units are built to study the modeling ability by both. These are compared against the baseline phone based unit selection voice.

Segmentation of the database in terms of graphemes and trigraphemes is automatically done using the EHMM labeller [8] in FESTVOX. For the grapheme based system, dictionary representation of each word is in terms of its graphemes, e.g. "# c a t #" for the word "cat". For the trigrapheme system, the context is also tagged to the representation, as in "#ca cat at#" for "cat". The acoustic models are trained using the iterative Expectation Maximization (EM) algorithm. Analyzing the likelihood with increasing iterations of training, it is noted that trigrapheme models converge faster and have a higher likelihood than their grapheme counterparts. This confirms that the context information helps improve the precision of the otherwise gross acoustic relevance of graphemes. Similarly, phone based models converge faster and have a higher likelihood than the trigrapheme models. After the training is complete, the utterances are segmented in terms of the units. Figure 2 shows sample segmentations of the phrase "Robbery, bribery, fraud" using grapheme, trigrapheme and phone units respectively.

After segmentation, the units are further clustered using context information for use during synthesis. In the unigrapheme and trigrapheme systems, the clustering for building the units is done with a context of size 2. This implies that for the unigrapheme based synthesizer, two neighboring unigraphemes on either side are given as the context. The trigrapheme system uses the two neighboring trigraphemes (effectively, 3 unigraphemes on either side) as the context. The splitting at the intermediate nodes is done on the basis of raw entropy. This is in contrast to the conventional approach of using higher level linguistic questions to capture the context. Thus the approaches presented in this paper bear minimal assumptions on knowledge of the language and hence are rapidly portable across languages.

Table 1. Number of units in each task

Synthesizer-Type	#Units
Phone	40
Unigrapheme	27
Trigrapheme	2984

To evaluate the performance of these synthesizers, a set

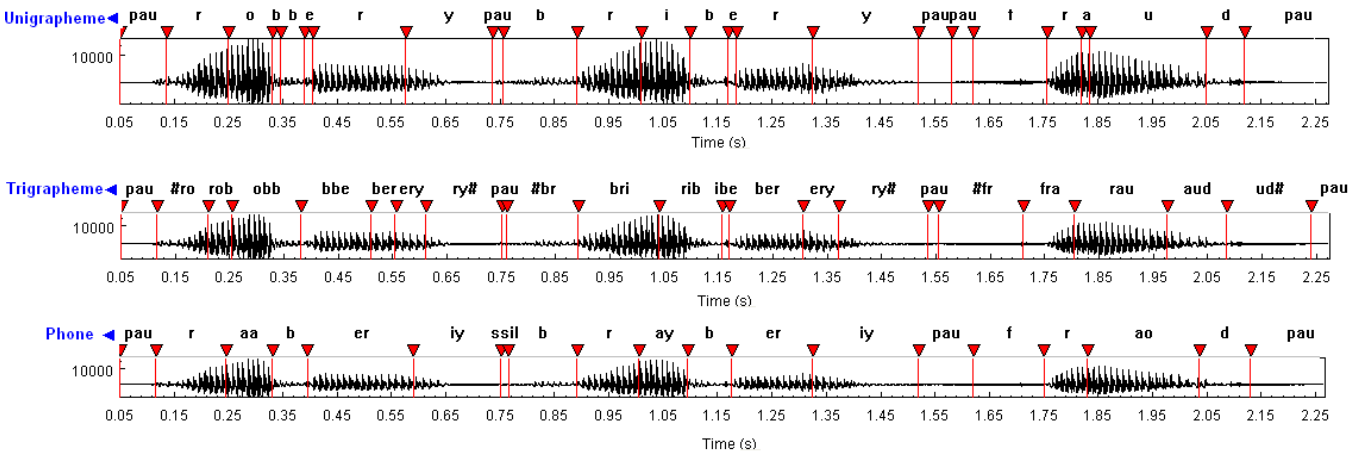

Fig. 2. Segmentations for the phrase "Robbery, bribery, fraud" using unigrapheme, trigrapheme and phone units.

of 15 sentences were synthesized. These 15 sentences were randomly chosen from Gutenberg text; however they aren't part of the ARCTIC dataset. Seven subjects were asked to listen to the sentences synthesized by phone, unigrapheme and trigrapheme based synthesis systems and were asked to score each sentence between 0 and 5 (0-worst and 5-best). The average scores of test sentences from different synthesizers are shown in Table 2. It could be observed the trigrapheme based speech synthesizer performs significantly better than unigrapheme based system. The scores obtained by trigrapheme based speech synthesis system are close to phone based speech system which shows that early tagging of contextual information plays a significant role in building speech synthesis systems.

5. TRIGRAPHEME BASED SPEECH RECOGNITION

State-of-the-art speech recognizers use context-dependent phonemes as the acoustic modeling units. The context dependence is incorporated at the time of state tying during training, where phone states belonging to 'similar' context share their parameters. The similarity is determined by a set of linguistic questions that decide on which states to cluster together. This dependence on existing knowledge from the language makes it suboptimal to use techniques across languages. This is the rationale behind investigations for other easily available modeling units. Attempts in the use of grapheme as a modeling unit for speech recognition have been reported in phonetic or partially phonetic languages. Most research on grapheme based recognition has focused on improving clustering to disambiguate various letter contexts during training. In this paper, we explore another dimension of introducing context.

The success of the trigrapheme units in speech synthesis (Sec. 4) has motivated us to evaluate them for applicability in recognition. In this work, we introduce context information for graphemes early on, at the choice of the modeling unit. An advantage of forcing context in this way is more discriminative modeling at the level of the units even before the application of conventional clustering routines for context disambiguation. Although, it may be argued that the technique is likely to face a data sparsity problem, it seems from our preliminary results that it is essential to disambiguate the otherwise gross grapheme units.

Table 2. Average ratings of test sentences from different synthesizers. The survey is taken by 7 subjects

S. No	Phoneme	Grapheme	Trigrapheme
1	3.71	2.57	2.86
2	3.79	2.43	3.79
3	3.86	2.29	3.36
4	2.71	2.71	3.29
5	3.29	2.71	4.50
6	3.36	2.57	3.00
7	2.93	3.71	3.79
8	4.07	2.86	3.71
9	3.29	2.14	3.21
10	2.93	2.71	2.93
11	4.07	2.57	3.00
12	3.86	3.93	3.71
13	3.79	2.29	3.21
14	3.79	3.14	3.36
15	3.14	3.14	3.64
Average	**3.50**	**2.79**	**3.42**

In order to evaluate the effectiveness of trigraphemes for speech recognition we built a trigrapheme decoder and compared the performance with that of unigrapheme and phone decoder. In these recognition experiments, each unit is represented by a 3 state context independent HMM with 2 Gaussians per state. Speaker specific acoustic models were built

on the RMS voice of ARCTIC dataset.

In all the three cases, Viterbi algorithm is used to decode through the search space of units. It has been well noted that decoding through an exhaustive search space is highly error-prone. The size of the space differs in the three systems with respect to the number of acoustic units each has and the length of the utterance to be decoded. Since trigrapheme units are higher in number the decoding took more time than the phone decoder or unigrapheme decoder. Due to the large number of classes in trigraphemes we expected more confusion among the units and was informally observed to be so in the decoded sequences of trigraphemes in the initial experiments. Moreover, we have used the same amount of limited data to train phone, unigrapheme and trigrapheme decoder.

The evaluation of these decoders was done by calculating the unit error rate as follows. Unit error rate takes into accounts insertions, deletions and substitutions similar to word error rate in speech recognition.

- Phone Recognizer: The decoded sequence of phones was compared against the expected sequence of phones
- Unigrapheme Recognizer: The decoded sequence of graphemes was compared against the expected sequence of unigraphemes.
- Trigrapheme Recognizer: Given the decoded sequence of trigraphemes, each trigrapheme was stripped off the tagged context. Thus the sequence of stripped-off trigraphemes was compared against the expected sequence of unigraphemes.

Table 3 shows the number of units in each decoder and the error rate of the three.

Table 3. Performance of exhaustive decoding by each recognizer

Modeling unit	#Units	Error Rate
Phone	40	55.3%
Unigrapheme	27	57.53%
Trigrapheme	2984	43.5%

The worse performance of unigrapheme decoding than phone decoding is as expected since unigraphemes seem to capture gross distributions and are far more ambiguous than phone models. It is also evident from the numbers that trigraphemes model the acoustics with a higher precision than even their phone counterparts. Though it is not investigated how well this holds in full scale decoding but they seem to be better descriptors of the acoustics.

6. CONCLUSIONS

In this paper, we have investigated the significance of early tagging of graphemes with contextual information and their effectiveness in building speech synthesis and speech recognition systems. From the perceptual evaluation tests, we have observed that performance of trigrapheme based synthesizer seem to be close to that of phone based synthesis system. From the error rates computed for the phone based and trigrapheme recognizers, it is also evident that early tagging of graphemes perform better than phone recognizer. While the concept of early tagging seems to be a simple trick in building speech synthesis and speech recognition systems, it seems to hold promise in building speech systems, specifically in the case of minority languages.

7. ACKNOWLEDGEMENT

The authors wish to thank Hussien Seid Worku, P. Lakshmi Narasimham, E. Veera Raghavendra, Sachin S. Joshi, Srinivas Desai, E. Uday Kumar and Venkatesh Keri for patiently evaluating the synthesizers.

8. REFERENCES

[1] M. Davel and E. Barnard, "The efficient generation of pronunciation dictionaries: machine learing factors during bootstrapping," in *ICSLP2004*, Jeju, Korea, 2004.

[2] A. Black and A. Font Llitjós, "Unit selection without a phoneme set," in *IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, CA., 2002.

[3] M. Killer, S. Stüker, and T. Schultz, "Grapheme based speech recognition," in *Eurospeech 2003*, Geneva, Switzerland., 2003.

[4] S. Hailemariam and K. Prahallad, "Extraction of linguistic information with the aid of acoustic data to build speech systems," in *Proc. of IEEE ICASSP*, Honolulu, USA., 2007.

[5] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Proceedings of EUROSPEECH'97*, 1997, pp. 601–604.

[6] J. Kominek and A. Black, "The CMU ARCTIC speech databases for speech synthesis research," 2003.

[7] A. Black and K. Lenzo, "FestVox: Building voices in the Festival Speech Synthesis System," `http://festvox.org/bsv/`, 2000.

[8] Kishore Prahallad, Alan Black, and Ravishankar Mosur, "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis," in *Proc. of IEEE ICASSP*, Toulouse, France, 2006.