IIIT Hyderabad's entry to Blizzard Challenge 2016

Sai Sirisha Rallabandi, Sai Krishna Rallabandi and Suryakanth V Gangashetty

Speech and Vision Laboratory
International Institute of Information Technology, Hyderabad, India
{sirisha.rallabandi,saikrishna.r}@research.iiit.ac.in, svg@iiit.ac.in

Abstract

In this paper, we are presenting IIIT-H's system, designed for the synthesis of storybooks as a part of the Blizzard Challenge 2016. We have extended unit selection and concatenation based system that was designed for the previous Blizzard Challenge 2015 by employing prosodic prediction module using a continuous representation of text. More specifically, we use a matrix factorization based approach to obtain dense representation at the phoneme, word level followed by a recurrent neural network based method to obtain dense representation at the sentence level. We use them to build a duration model with an intention to capture the variations in prosody due to the nature of children's story books. We have also investigated the use of sentence level vectors for modeling prosody.

Index Terms: Speech Synthesis, Blizzard Challenge, Recurrent Neural Networks

1. Introduction

This is our second entry to Blizzard Challenge. The submission we made in the year 2015 was a unit selection and concatenation based one where the selection of syllable based units was performed using a prosodic matching function on preclustered syllable level units. The selected units were concatenated using an overlap method which aims at maximizing the continuity at the point of join. The task in this year's challenge was to generate children's stories, which have an interesting characteristic that they have embedded emotions within them. Therefore, the objective is not just to synthesize the input text but also to generate it in such a way that a human presents it, using styling and phrase breaks.

One of the important attributes of a speech signal which contributes to the styling is its prosody and variations in it. Prosody is usually modeled using statistical methods such as decision trees, random forests, hidden markov models and deep neural networks, etc. The idea is to use a statistical model to map the input text to the output speech parameters such as duration and fundamental frequency. The input text is typically represented in the form of a one-hot-k vector and the output is normalized. In recent years, there has also been work towards using continuous representation of text as the input, with an intention that this helps the deep neural network based models get a better generalization while modeling prosody.

For the current submission, we adopted this approach and extended our system to a statistically guided unit selection and concatenation based system. Specifically, we have made the following extensions:

- We have used matrix factorization based approach to obtain continuous representation of text at the phone and word levels. These representations were then used to model the prosodic parameters.
- We have used language model configuration of recurrent neural network to obtain continuous representation of text at the sentence level and used it in the prosodic models.
- We have employed a signal cross correlation based continuity measure at the viterbi selection of units.

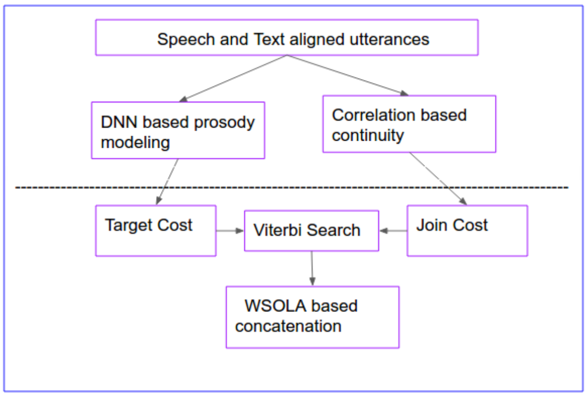
Speech and Text aligned utterances
DNN based prosody modeling
Correlation based continuity
Target Cost
Viterbi Search
Join Cost
WSOLA based concatenation

Figure 1: *Overview of the Hybrid Speech Synthesis System*

This paper is organized as follows. Section 2 and 3 describe the new methods that have been applied in our system. Section 4 gives some detailed system building. The discussions and conclusions are in section 5 and 6.

2. System Description

2.1. Data

The database used in the challenge was provided by Usborne Publishing Ltd. and consists of the speech and corresponsing text of fifty children's audiobooks spoken by a native british speaker. We are given about 5 hours of speech data which includes approximately 2 hours of pilot data from last years Blizzard Challenge. We have made use of segmentation details made available by another participant and removed the 'bell' sounds which were present in the speech and all the other expressions like 'uh..', 'hm..'. The total duration of the audio is approximately 4.5 hours after segmentation. Two audiobooks were held out to act as an development set to fine tune the prosodic models.

2.2. Pronunciation for out of lexicon words

There were around 130 words in the training corpus which were absent from the CMU pronunciation dictionary. We have used word to phone mapping using automatic epsilon scattering method to obtain the pronunciations for those words. In this method, each letter is assumed to be specifying a phonetic correspondence to one or more phones. If the letters are not mapped to a phone then epsilon is used. As a fixed sized learning vector is required to build a model for learning word-phone mapping rules, we need to align the letter (graphemic) and phone sequences. For this automatic epsilon scattering method [14] was employed, where the central idea is to estimate the probabilities for one letter (grapheme) G to match with one phone P, and then use string alignment to introduce epsilons maximizing the probability of the alignment path of that word. Once all the words have been aligned, the association probability is calculated again and so on until convergence. Once the alignment between the each word and the corresponding phone sequence was complete, we built the phone model using random forests [15].

2.3. Selection of Sentences for Modeling Prosody

In the context of expressive synthesis systems which are used to generate text such as the current task, selection of appropriate sentences for building the statistical models is necessary. This data pruning involves removal of spurious units (which may be a result of mislabeling or bad acoustics) and units that are redundant in terms of prosodic and phonetic features. It was shown that pruning spurious units improves TTS output [1, 2, 3, 4, 5, 6] while pruning redundant units reduces database size thus enabling portability [7, 8, 9] and real-time concatenative synthesis [10, 11, 12]. We have used a clustering method similar to the one in [1] where each unit is represented as a sequence of MFCC vectors, and clustering using decision tree proceeds based on questions related to prosodic and phonetic context. Each unit is then assessed for its frame-based distance to cluster center. Units which lie far from their cluster centers are termed as outliers and hence pruned. We used unit duration [6] obtained from the ASR system as confidence measures.

2.4. *Prosodic Matching Criterion*

The intuition behind the prosodic matching criterion is that a unit is perceived better if its successor and predecessor has similar *prosodic behaviour* which can algebraically be calculated based on a loss function. A rudimentary way to define such a loss function can be based on the features of the individual units and can be calculated using

$$m(loss) = \sum_{1}^{n} \left| F_n^{k-1} - F_n^k \right| \qquad (1)$$

where n is the number of feature. The features typically employed in the loss function are F0, log energy, etc. During concatenation, a manifestation of k^{th} phone (k = 2.3..n) where n is the number of syllables) is selected if it follows:

- Phonetic Context - The k^{th} phone should have the coarticulative continuity of the last phone of the $(k-1)^{th}$ phone.
- Similarity to the $(k-1)^{th}$ phone as per the prosodic criterion.

A special case of this approach is the selection of first phone which is done based on the next phone. When using phones as the basic units for Indian languages, it was shown that it is enough to consider just the four frames at the boundary for achieving a good quality concatenation [16]. We have used the same intuition in the design of the current system. Therefore, the loss function can be represented as:

$$m(loss) = \sum_{1}^{4} \sum_{1}^{n} \left| F_n^{k-1} - F_n^k \right| \qquad (2)$$

The features used in the function are discussed below:

- *Target Cost* - As the basic units were already clustered based on their phonetic positions, a very simple way to account for the difference between in terms of target cost would be based on the duration of units. The mean duration for each of the units is computed using all the occurrences in the database. Thus, the units with minimum distance from this mean value have a higher probability in getting selected when the total cost is obtained.
- *Join cost* - Traditionally prosodic features such as duration and f0 are used to calculate acoustic similarity between two units with provisions to use the cepstral features mel frequency cepstral coefficients (MFCC) in the vector. Preliminary informal analysis was performed to see the influence of addition of energy and second formant(F2) in the vector. The intuition behind using F2 was based on the studies from locus equations that showed its influence on the determination of place of articulation and extent of coarticulation. However, as significant improvements were not noticed, it was not used in the final feature set to keep the vector compact. The final features that were used were MFCCs, energy and F0.

2.5. Viterbi search

The equation 3 below explains the way the total cost is computed. The term *Tdist(U_i)* is the difference between duration of unit U_i and the predicted duration, and the term $Jdist(Ui, Ui-1)$ is the join cost of the optimal coupling point between candidate unit U_i and the previous candidate unit it is to be joined to. W_1 and W_2 denote the weights given to target and join costs respectively. N denotes the number of units to be concatenated to synthesize the sentence in question. We then used a Viterbi search to find the optimal path through candidate units that minimized the total cost which is the sum total of target and concatenation costs.

$$\text{Total cost} = \sum_{i=1}^{N} W_1 Tdist(U_i) + W_2 Jdist(U_i, U_{i-1}) \qquad (3)$$

2.6. *Waveform Concatenation*

Eventhough the prosodic function generates the selection of units which have minimum distance in the feature space, there is no guarantee that these units result in a smooth speech when concatenated. Therefore, typically some sort of smoothing function is applied while concatenating the selected units. For this, a variation of Waveform Similarity Overlap Addition (WSOLA)[17] method was formulated. Specifically, the algorithm was reformulated in order to first find a suitable temporal point for concatenating the units at the boundary. This is done so that the concatenation is performed at a point where maximal similarity exists between the units. In other words, we tried to ensure that sufficient signal continuity exists at the concatenation point. There are two ways this can be achieved: Maximizing the cross correlation between the segments or minimizing the Average Magnitude Difference between the segments. For the current design, we have used the cross correlation based method based on an informal evaluation of both the methods. Then, the units are concatenated at the point of best correlation using crossfade technique[18] to further remove the phase discontinuities.

3. Continuous Representation of Text

In this section, we describe the training procedure used to obtain the continuous representation at various levels for the the input text. We have used deep neural networks as the statistical framework to model the durations of the selected triphones.

3.1. Continuous Representation at the phone and word Level

Recently, there has been a lot of work supporting the representation of words as dense vectors, derived using various training methods inspired from neural-network language modeling [19][20]. Such distributed representations, also termed as vector space models (VSMs) have previously been applied to Speech synthesis from text in [21], where prediction models are built at various levels of analysis (letter, word and utterance) from unlabelled text. We have derived the distributed representation at the phone level similar to the approach used in SkipGram Model[20]. Inspired by [22], we pose the task as a matrix factorization problem and solve it using Symmetric Singular Value Decomposition. To build these models, co-occurrence statistics are gathered in the form of matrix to produce high-dimensional representations of the distributional behaviour of the chosen unit in the corpus. Appropriate lower dimensional representations are obtained by approximately factorising the matrix of raw co-occurrence counts by the application of Singular Value Decomposition(SVD). Further details about the exact procedure to obtain the representations can be obtained from [23].

3.2. Continuous Representation at the sentence Level

As prosody is a suprasegmental feature, it might be better to model its variations at the sentence level rather than at the phone or word level. For the current submission, we have used a recurrent neural network language model to obtain the sentence level vectors. In a nutshell, the network consists of three layers: input, hidden and output layer. The input layer represents the current word using 1-of-N coding and encodes it, the hidden layer encodes the current sentence upto the current word and the output layer predicts the probability of the next word. As our intention is to obtain the representation from the hidden layer, we have factorized the output layer using class based information. The network is trained using back-propagation through time (BPTT), an extension of the back-propagation algorithm for recurrent neural networks.

3.3. Correlation based Continuity Measure

Although unit selection based approach generally synthesizes speech with high-level of intelligibility and naturalness, it is bothered by the stability problem that critical errors will occasionally occur and ruin the perception of the whole utterance.In our submission, we try to address this issue and the motivation for our approach comes from the understanding that the human speech database is fully comprised of naturally evolving adjacent speech frames, forming sequences of audibly perfect joins. The adjacent speech frames are highly correlated with each other. We try to emulate this correlated behavior in our synthesis framework. Specifically, we investigate the use of a continuity metric targeted at maximizing such correlation between the units during synthesis. There are two ways of using the correlation between the units in a unit selection synthesis framework. One way is use the knowledge of signal correlation during the concatenation of the selected units so that they are joined at the point of maximum correlation between the units as mentioned in [24]. The other way, which we focus on in the current submission, is to directly use the correlation as a sub cost in the join cost, thereby controlling the selection of the units themselves. We have employed two formulations for estimating correlation between the units: Cross correlation based formulation and Average Magnitude Difference Function (AMDF) based formulation.

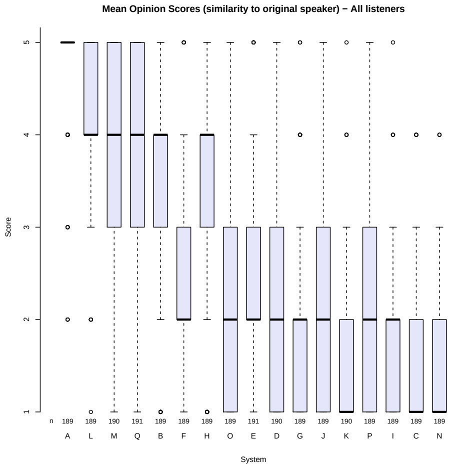

Figure 2: *MOS scores from all the listeners for similarity.*

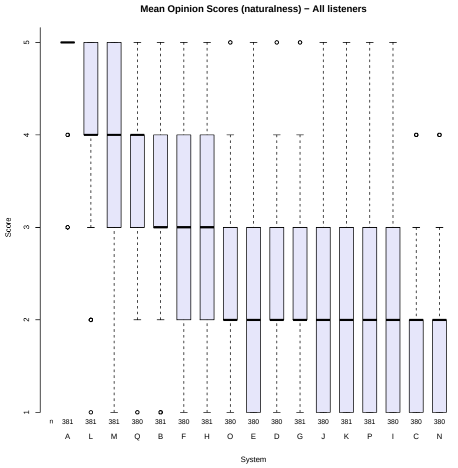

Figure 3: *MOS scores from all the listeners for naturality.*

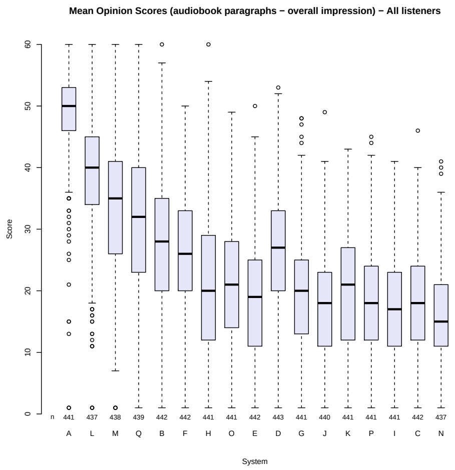

Figure 4: *MOS scores from all the listeners for the overall performance.*

4. Evaluation

The evaluation was conducted under various categories: pleasantness, speech pauses, stress, intonation, emotion, listening effort. Mean opinion scores of our system as provided by all the listeners are depicted in the figures 2 and 3, and the overall performance of the system is shown in the figure 4. The identifier of our system is E.

5. Conclusion

To summarize, for the Blizzard Challenge 2016 we have developed a hybrid system, where we have exploited the continuous representation of input text in the form of phones and words using matrix factorisation method and sentence level representation by RNNLMs. Then we have employed the viterbi search extending our previous work by embedding the correlation based continuity metric for the selection of the appropriate units. Concatenation of the obtained units was done using WSOLA.

6. References

[1] A. W. Black and P. A. Taylor, “Automatically clustering similar units for unit selection in speech synthesis.” 1997.

[2] H. Hon, A. Acero, X. Huang, J. Liu, and M. Plumpe, “Automatic generation of synthesis units for trainable text-to-speech systems,” in *Proc. ICASSP*, vol. 1. IEEE, 1998, pp. 293–296.

[3] H. Lu, Z. Ling, S. Wei, L. Dai, and R. Wang, “Automatic error detection for unit selection speech synthesis using log likelihood ratio based SVM classifier.” in *Proc. Interspeech*, vol. 10, 2010, pp. 162–165.

[4] J. Adell, P. D. Agüero, and A. Bonafonte, “Database pruning for unsupervised building of text-to-speech voices,” in *Proc. ICASSP*, vol. 1, 2006, pp. I–I.

[5] L. Wang, Y. Zhao, M. Chu, F. K. Soong, and Z. Cao, “Phonetic transcription verification with generalized posterior probability.” in *Proc. Interspeech*, 2005.

[6] J. Kominek and A. W. Black, “Impact of durational outlier removal from unit selection catalogs,” in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[7] H. Lu, W. Zhang, X. Shao, Q. Zhou, W. Lei, H. Zhou, and A. Breen, “Pruning Redundant Synthesis Units Based on Static and Delta Unit Appearance Frequency,” in *Proc. Interspeech*, 2015.

[8] R. Kumar and S. P. Kishore, “Automatic pruning of unit selection speech databases for synthesis without loss of naturalness.” in *Proc. Interspeech*, 2004.

[9] V. Raghavendra and K. Prahallad, “Database pruning for Indian language unit selection synthesizers,” pp. 67–74, 2009.

[10] D. Schwarz, G. Beller, B. Verbrugghe, S. Britton *et al.*, “Real-time corpus-based concatenative synthesis with catart,” in *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx), Montreal, Canada*. Citeseer, 2006, pp. 279–282.

[11] R. E. Donovan, “Segment pre-selection in decision-tree based speech synthesis systems,” in *Proc. ICASSP*, 2000.

[12] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996. ICASSP-96, 1996*, vol. 1. IEEE, 1996, pp. 373–376.

[13] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, “Confidence measures for large vocabulary continuous speech recognition,” *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 3, pp. 288–298, 2001.

[14] V. Pagel, K. Lenzo, and A. Black, “Letter to sound rules for accented lexicon compression,” *arXiv preprint cmp-lg/9808010*, 1998.

[15] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] V. R. Lakkavalli, P. Arulmozhi, and A. Ramakrishnan, “Continuity metric for unit selection based text-to-speech synthesis,” in *Signal Processing and Communications (SPCOM), 2010 International Conference on*. IEEE, 2010, pp. 1–5.

[17] W. Verhelst and M. Roelands, “An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech,” in *Proc. ICASSP*, vol. 2, 1993, pp. 554–557.

[18] T. Hirai and S. Tenpaku, “Using 5 ms segments in concatenative speech synthesis,” in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[19] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, “Recurrent neural network based language model.” in *Proceedings of INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, 2010, pp. 1045–1048.

[20] T. Mikolov, S. Kombrink, L. Burget, J. H. Cernocky, and S. Khudanpur, “Extensions of recurrent neural network language model,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011*. IEEE, 2011, pp. 5528–5531.

[21] O. Watts, “Unsupervised learning for text-to-speech synthesis,” Ph.D. dissertation, University of Edinburgh, 2012.

[22] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *Advances in Neural Information Processing Systems*, 2014, pp. 2177–2185.

[23] S. K. Rallabandi, S. S. Rallabandi, P. Bandi, and S. V. Gangashetty, "Learning continuous representation of text for phone duration modeling in statistical parametric speech synthesis," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 111–115.

[24] S. K. Rallabandi, A. Vadapalli, S. Achanta, and S. V. Gangashetty, "Iiit-h's entry to blizzard challenge 2015," in *proceedings of Blizzard Challenge Workshop 2015*.