International Journal of Engineering Research in Electronic and Communication Engineering (IJERECE) Vol 3, Issue 5, May 2016

# Correlation Based Continuity Metric for Concatenative Speech Synthesis

[1] Naina Teertha, [2] Dr.R.Kumaraswamy[3] Sai Sirisha Rallabandi, [2] Sai Krishna Rallabandi
[5] Venkatesh Potluri [6] Suryakanth V Gangashetty
[1][2] Speech and Vision Laboratory, International Institute of Information Technology, Hyderabad, India
[3][4][5][6]Siddaganga Institute of Technology, Tumkur, Karnataka, India
[1] nainai91i98@gmail.com [2] sirisha.rallabandi@research.iiit.ac.in, [3]saikrishna.r@research.iiit.ac.in
[4] hyrkswamy@gmail.com [5] svg@iiit.ac.in

***Abstract*— In this paper, we investigate the use of a new continuity measure based on maximum signal correlation for optimal selection of units in concatenative text-to-speech (TTS) synthesis framework. We explore two formulations for calculating the signal correlation: cross correlation (CC) based and average magnitude difference function (AMDF) based. We first perform an initial experiment to understand the significance of the approach and then build 5 experimental systems which are available a web demo. Evaluations on 30 sentences each for Telugu and Hindi by native users of the languages show that the proposed continuity measure results in more natural sounding synthesis.**

***Index Terms*: Unit selection speech synthesis, Target cost, Join cost, Cross correlation, Average magnitude difference function.**

## I. INTRODUCTION

A children story can be divided into three parts: (i) introduc- tion, (ii) main and (iii) climax. Introduction part comprises of introducing characters that are involved in the rest of story. It also describes time and place of the event. Main part constitutes the core component of the story. It has series of events and ac- tions that relate to a central theme of the story. The central idea of the story and moral are described in the climax part. In this work, we are classifying children stories into three story gen- res namely fable, folk-tale and legend based on their structure. Fable is a short tale involving animals as essential characters. Folk-tale is a traditional story that is passed on in spoken form from one generation to the next. Legend is a semi-true story carrying significant meaning or symbolism for the culture in which it originates. It is based on historical factors of a particu- lar geographic region.The basic goal of this work is to develop a story speech synthesis system. Given a story text, the system should synthesize speech as narrated like a story teller. A story- teller narrates a story by varying prosody like pitch, duration and intensity. It is also observed that narration style depends on the story genre and hence there is a need to identify story genre from the given story text for story synthesis. Given a neutral Text-to-speech (TTS) system and prosody modification rules associated with each story genre, the TTS system should synthesize story speech. Recently, syllable-based unit selection neutral TTS systems were developed in 13 Indian languages [1]. We need to integrate prosody modification and text processing modules for the existing neutral TTS systems to synthesize the desired story speech. With this motivation, we tried to explore story genre classification and how story genre information is embedded in different parts of the story for Hindi and Telugu. In this work, we have manually divided the stories based on their structure. We proposed a new feature based on Parts-of-Speech (POS) for story classification. The motivation for considering the POS information for story classification is supported by the observations like more named entities in stories and importance of POS tags like nouns, adjectives, quantifiers for discriminat- ing between story genres. From the literature, it has been ob- served that there is no existing works on story classification for Indian languages and features related to POS have not been ex- plored for story classification. Hence, in this study we are us- ing POS features for story classification in addition to keyword based features.

## II. RELATED WORK

Within the context of a story telling TTS application, a per- ceptual study to identify emotions in children stories was car- ried out [2]. In [3], children stories have been analyzed for identification of characters and personality attributes of char- acter like age and gender. Story classification is a typical doc- ument classification problem, and it has been carried out for different domains using different approaches across languages. Most widely used approaches for text

classification for Indian languages are WordNet and Machine Learning approach. In [4], ontology and hybrid based approach for classification of Punjabi text documents was proposed. They developed sports specific ontology for Punjabi and prepared gazetteer lists such as middle names, last names, abbreviations etc., for Named Entity Recognition task. In [5], Marathi articles were classi- fied using different classifiers and built rule based stemmer and Marathi word dictionary to reduce the dimensionality of fea- ture vectors. In [6], Kannada web pages were classified us- ing various pre-processing agents. Pre-processing steps like language identification, sentence boundary detection, stemming and stopword removal are applied on the webpage content be- fore classification. In [7], stop words and restrictions based on word occurrence were used for dimensionality reduction and classified manually collected Kannada sentences from Kannada Wikipedia. In [8], Tamil documents were classified using Arti- ficial Neural Network (ANN) and Vector Space Model (VSM). Their experiments concluded that ANN is better for more rep- resentative collection and captured the non-linear relationships between the input document vectors and the document cate- gories than that of VSM. In [9], Telugu news articles were clas sified into four categories: Politics, Sports, Business and Cin- ema using NB classifier. In [10], language independent, corpus- based machine learning techniques were used for text catego- rization in ten major Indian Languages. But, there is no existing work on story classification for Indian languages which led to the motivation of the present study.

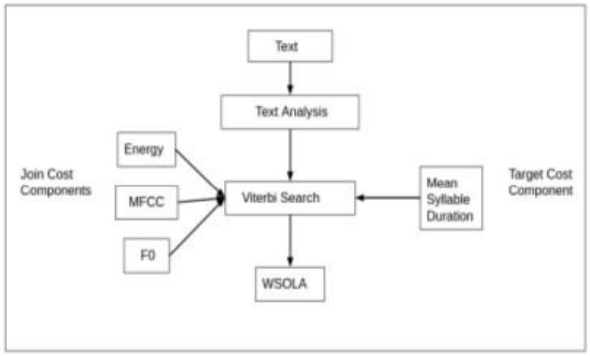


**Figure 1:** ***Overview of the Framework.***

## III. OVERVIEW OF THE CURRENT FRAMEWORK: BASELINE SYSTEM

In this section, we give a brief overview of our synthesis system, as used in [1]. The framework follows a frontend/back-end ar- chitecture with Natural Language processing as front end and Digital signal processing module as back end.

### *2.1. Unit Size*

Earlier work on Indian languages [2] suggested that a syllable based approach to synthesis could lead to more reliable qual- ity. There are a number of reasons for this, some of them being that (a) the syllable units can capture coarticulation better than phonemes, (b) the number of concatenation points decreases when syllable is used as the basic unit, (c) syllable boundaries are characterized by regions of low energy and therefore audi- ble mismatches at the boundary are hardly perceived, etc. Also in the context of Indian languages, the number of polysyllablic words is huge. Due to these advantages, we have chosen sylla- ble as the basic unit in our concatenative synthesis framework. The syllable in Indian languages is of the form V, VC*, C*V and C*VC* where V is the vowel and C is the consonant.

### *2.2. Steps to build voices*

#### *2.2.1. Automatic Alignment of Speech and Text*

For segmenting the audio data we used the procedure described in [3] which is based on an HMM forced alignment algorithm. The alignment has been performed without any change or super- vision as it closely developed to the TTS front-end component.

#### *2.2.2. Preclustering the units*

It was seen that syllables of the same type can be easily dif- ferentiated depending on their position in the word [4]. In ad- dition, syllables occurring at the beginning of the word are of longer duration than the syllables occurring at the middle and end of a word [5], [6]. The energy and pitch were also found to vary depending on the position of the syllable in the word [7]. Therefore, weve performed pre-clustering based on position of the syllable in a word, i.e syllables of the same type were pre- clustered as begin, middle and end by appropriately depending on their position in the word, in the original context. In case a syllable of appropriate position is not available during synthe- sis, an order of preference is used to pick a syllable of the same type occurring at an alternate position. During synthesis, if the required begin or an end syllable is not present in the database, middle syllable is preferred. If the required middle syllable is not present, a syllable from a word beginning is selected instead.

#### *2.2.3. Target and Join costs*

Typically mel frequency cepstral coefficients (MFCC) are used to calculate distance between two units accompanied by dura- tion and F0 of the unit. Preliminary analysis on the data showed that the energy of units play a major role in syllable unit syn- thesis. We have therefore included log energy, MFCC, dynamic features of MFCC ( deltas and double deltas), F0 and unit dura- tions as the acoustic features. We employed a target cost based on the distance from the mean duration of the syllables in the current version of the framework, following [8]. The mean du- ration 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. The join cost consists of the sub costs arising from log energy,

spectral and pitch based features. We follow the formulation similar to the one proposed in [9], [10] and calculate the spectral, f0 and energy based continuity metrics using 1, 4 and 2 boundary frames respectively[1]. The weights of the individual sub costs have been optimized manu- ally over a held out set from the training data.

#### *2.2.4. Concatenation based on waveform similarity*

We use an overlap addition based approach for smoothing the join at the concatenation boundaries. We first find a suitable temporal point for joining 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 try to en- sure that sufficient signal continuity exists at the concatenation point. For this, we use the cross correlation between the units as a measure of similarity between the units. Then, the units are concatenated at the point of best correlation using crossfade technique[11] to further remove the phase discontinuities. The number of frames used to calculate the correlation is limited by the duration of the available subword unit. In the current frame- work, we have used the last two frames of the individual units to calculate the cross correlation. We have used reduced vowel epenthesis based backoff [12] strategy to synthesize the miss- ing units and word to native speaker phone mapping [13] for the English words.

## III. EXPERIMENTS

### *3.1. Data*

We have used the database provided as a part of Blizzard Challenge 2015 for the purpose of the current investigation. Although the data was released for 6 languages, we have used Telugu( a Dravidian language) and Hindi( an Indo-Aryan language) databases for our experiments, promarily as they are from different language families. The other reasons for selecting these languages were the availability of native speakers for testing and larger database size (4 hours) compared to the other languages (2 hours). The training and the test set have been used as is except for leaving out 15 training sentences as a held out set to validate the findings and tune the weighting functions.

***Table 1: Preference Test on Telugu. The percentages are shown for each system.***

| Cross Correlation | 4 F vs 2 F | 2 F vs 1 F | 4 F vs 1 F | Average Magnitude Difference | 4 F vs 2 F | 2 F vs 1 F | 4 F vs 1 F |
|---|---|---|---|---|---|---|---|
| Prefer 4 F | 23 | - | **47** | Prefer 4 F | 29 | - | **46** |
| Prefer 2 F | **51** | **73** | - | Prefer 2 F | **46** | **79** | - |
| Prefer 1 F | - | 7 | 21 | Prefer 1 F | - | 9 | 33 |
| No Preference | 26 | 20 | 32 | No Preference | 25 | 12 | 21 |

***Table 2: Preference Test on Hindi. The percentages are shown for each system.***

| Cross Correlation | 4 F vs 2 F | 2 F vs 1 F | 4 F vs 1 F | Average Magnitude Difference | 4 F vs 2 F | 2 F vs 1 F | 4 F vs 1 F |
|---|---|---|---|---|---|---|---|
| Prefer 4 F | 22 | - | **42** | Prefer 4 F | 22 | - | **44** |
| Prefer 2 F | **53** | 77 | - | Prefer 2 F | **59** | **73** | - |
| Prefer 1 F | - | 13 | 23 | Prefer 1 F | - | 17 | 30 |
| No Preference | 25 | 10 | 35 | No Preference | 19 | 10 | 26 |

### *3.2. Incorporation of Correlation in Join Cost*

Correlation between the units can be incorporated as one of the subcosts in the join cost thus affecting the selection of the units. However, it has to be validated that the selected units are correlated with each other and increase the naturalness of the synthesized signal. An initial experiment was designed to understand if the synthesis using correlation as subcost calculated from few frames of the units is acceptable perceptually. A set of 30 words in both the languages were selected. This set included 4 missing syllable units and 3 borrowed words (English words). Each of these words was synthesized by incorporating correlation as the sub cost in the join cost, with the correlation calculated using cross correlation based formulation(15 words) and Average Magnitude Difference Function(15 words) based formulation. For each word, 3 files were synthesized, using 4,2 and 1 frames at the boundary to calculate the correlation between the units. It is important to note that the correlation score obtained us- ing cross correlation formulation has to be maximized where as the score obtianed via AMDF has to be minimized. Forced preference test was performed by native speakers of both the languages. We have followed the same procedure mentioned in [28] and the results are summarized in tables 1 and 2. The results indicate that the correlation based approach is indeed preferred by the users, in both the languages and for both the formulations. They also indicate that for both the formulations, using 2 frames to the left and right at the boundary to calculate the correlation has received maximum preference.

### *3.3. Experimental Systems*

In this sub-section, we describe the experimental systems de- signed. Based on the inferences from the preference test, we have used 2 boundary frames to calculate the continuity metric in all of the experimental systems.

#### *3.3.1. Type A Systems - Systems CC and AMDF*

These are the experimental systems built only using cross- correlation and average magnitude difference function as the join cost components. In other words, system CC has only the cross correlation based continuity metric as the join cost and system AMDF has only the average magnitude difference function based continuity metric as the join cost. The intention behind building these systems is to understand if ensuring tem- poral correlation in the signal alone would suffice as the join cost to produce highly natural speech.

#### *3.3.2. Type B Systems - System Baseline + CC and System Baseline + AMDF*

These systems have been built to investigate the performance using the continuity measures in combination with the other subcosts. The weighting functions for each of the sub costs were optimized manually using a held out set of 15 utterances from the training set in both the languages.

### *3.3.3. Hybrid System - System Baseline + CC + AMDF*

This system has been built to see if the combination of both the formulations in addition to the existing sub costs achieves better performance than the individual systems.

### *3.4. Subjective Evaluation*

In order to evaluate the systems, we have performed subjective evaluations using procedure similar to Blizzard challenge lis- tening tests[1-6]. We have used the first 30 sentences from the test sets in both the languages for the listening tests. A set of 15 participants were made to listen to the synthesized files and rate the naturalness on a scale of 1 to 5, with 5 being the most natural and 1 being the most unnatural. The results from the listening test are shown in the figure 2. Type A systems have performed worser compared to the baseline system in both the languages and both the formulations, showing that the other sub costs have a significant role in the join cost. In line with our hypothesis, type B systems have better MOS scores compared to the baseline system, in both the languages and boht the formulations. Further, it is clearly evident that the hybrid system, using both the formulations has significantly outperformed the baseline system.

### *3.5. Conclusions*

In this paper, we performed an experimental analysis on the us- age of signal correlation as join cost in concatenative speech synthesis framework. As answers to the questions posed in section 1, we have observed that the continuity measures do make a perceptual difference and therefore serves as an impor- tant feature to be considered for obtaining more natural syn- thesis. We also found from the preference test that using 2 time frames for calculation of the correlation was preferred for both the formulations.However, when the measures were used in isolation(type-A systems), their performance is not very encouraging. When they are combined with the other costs(type-B systems and hybrid system), they outperform the baseline. Results on systems developed for the Telugu and Hindi languages provide evidence on the effectiveness of the proposed method.The samples and listening test results used in the experiments are available online via this link: https://goo.gl/XJgOUc

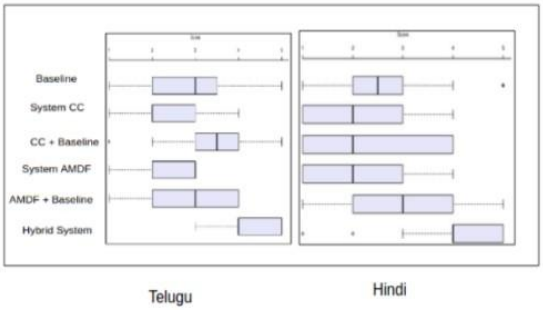


**Figure 2:** ***Results from the Subjective Evaluation. Figure de- picts box plots plotted based on the Mean Opinion Scores on a scale of 5.***

## REFERENCES

[1] S. K. Rallabandi, A. Vadapalli, S. Achanta, and S. V. Gangashetty, ‘‘Iiit-h’s entry to blizzard challenge 2015,’’ in *Interspeech 2015*.

[2] S. Kishore and A. W. Black, “Unit size in unit selection speech synthesis,’’ in *Eighth European Conference on Speech Communication and Technology*, 2003.

[3] K. Prahallad, A. R. Toth, and A. W. Black, ‘‘Automatic building of synthetic voices from large multi-paragraph speech databases.’’ in *INTERSPEECH*, 2007, pp. 2901–2904.

[4] H. A. Murthy, “Methods for improving the quality of syllable based speech synthesis,’’ 2008.

[5] M. V. Vinodh, A. Bellur, K. B. Narayan, D. M. Thakare, A. Susan, N. M. Suthakar, and H. A. Murthy, “Using polysyllabic units for text to speech synthesis in indian languages,’’ in *Communications (NCC), 2010 National Con- ference on*, Jan 2010, pp. 1–5.

[6] K. S. Rao and B. Yegnanarayana, “Modeling durations of syllables using neural networks,’’ *Computer Speech & Language*, vol. 21, no. 2, pp. 282–295, 2007.

[7] A. Bellur, K. B. Narayan, K. R. Krishnan, and H. A. Murthy, “Prosody modeling for syllable-based concatena- tive speech synthesis of hindi and tamil,’’ in *Communica- tions (NCC), 2011 National Conference on*, Jan 2011, pp. 1–5.

[8] H. R. Shiva Kumar, J. K. Ashwini, B. S. R. Rajaram, and A. G. Ramakrishnan, “Mile tts for tamil and kannada for blizzard challenge 2013,’’ in *Blizzard Challenge 2013 workshop, Barcelona, Catalonia*. CMU, 2013.

[9] V. R. Lakkavalli, P. Arulmozhi, and A. G. Ramakrishnan, “Continuity metric for unit selection based text-to-speech synthesis,’’ in *Signal Processing and Communica tions (SPCOM), 2010 International Conference on*, July 2010, pp. 1–5.

[10] S. K. H. Rajaram, BSR and A. Ramakrishnan, “Mile tts for tamil for blizzard challenge 2014,’’ in *Signal Process-*

*ing and Communications (SPCOM), 2010 International Conference on*. IEEE, 2010, pp. 1–5.

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

[12] V. Peddinti and K. Prahallad, "Significance of vowel epenthesis in telugu text-to-speech synthesis," in *Acous- tics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 5348–5351.

[13] N. K. Elluru, A. Vadapalli, R. Elluru, H. Murthy, and K. Prahallad, "Is word-to-phone mapping better than phone-phone mapping for handling english words?" in *ACL (2)*, 2013, pp. 196–200.