Proc. of Int. Conf. on Current Trends in Eng., Science and Technology, ICCTEST

Synthesis of Kannada Isolated Consonant-Vowel Coarticulations using Formant Synthesizer

Alfred Vivek D'Souza[1] and D.J Ravi[2]
[1]Assistant Professor, Vidyavardhaka College of Engineering, Mysuru,
Visvesvaraya Technological University, Belagavi
Email: alfreddsouza@vvce.ac.in
[2]Professor, Vidyavardhaka College of Engineering, Mysuru,
Visvesvarya Technological University, Belagavi
Email: ravidj@vvce.ac.in

***Abstract*—Formant analysis of various sound units play an important role in developing a functional Formant Synthesizer. This paper presents the result of formant analysis made on 15 different consonants occuring in context of 10 vowels and 2 diphthongs of Kannada language. The results were used to generate isolated Consonant-Vowel Coarticulations using Klatt like Formant Synthesizer.**

***Index Terms*— Coarticulations, Klatt Formant Synthesizer and Kannda CV synthesis.**

I. INTRODUCTION

Considerable efforts are being made since the early days of speech processing in understanding the generation of speech. Speech Synthesis is the branch of speech processing which deals with generation of human speech artificially using machines, known as Speech Synthesizers. Human vocal system generates a series of sounds in specific order to convey information. A Speech Synthesizer which aims to generate intelligible speech must have the capability of producing different types of sounds the human vocal system can produce.

The number and types of sounds that are produced when speaking, differs from one language to another. Sound units such as vowels and diphthongs can be pronounced independently where as consonats are often pronounced in conjunction with a vowel or diphthong, forming coarticulations. Coarticulation refers to the process of transition from one utterence to another. Coarticulations cause the vocal tract to vary its parameter in a complex way as opposed to the same sound units successively uttered in isolation. The goal of this study is to analyze how the vocal tract parameters change its configuation during coarticulations of Kannada language and synthesize such coarticulation using Formant Synthesizer. Kannada is one of the widely spoken languages of southern India having over 50 million speakers and has a rich set of sounds.

A. Speech Synthesizers and Current Trends in Speech Synthesis

Speech Synthesizers form the core of Text-To-Speech(TTS) systems and have variety of applications such as screen readers, that reads out the text on computer screens for visually impaired, talking help for those who

Grenze ID: 02.ICCTEST.2017.1.4

cannot speak due to physical inability, learning tool to know correct prounciation of words and in Natural Language Processing (NLP) for Human Computer Interface(HCI) to name a few. There are various techniques of speech synthesis few important ones are
Concatenative syntheisis, synthesis by analysis and syntheis, articulatory synthesis and formant synthesis.
Reference [1] discusses a Kannada Language TTS developed by MILE Lab, IISc , Bengaluru. The system is successfully developed using waveform concatenation approach. Work is done in the areas of selecting optimum units for concatenation and reducing the size of database.
Articulatory speech synthesis methods are considered complete solution to speech synthesis problem. It is possible to generate any utterance with required intonation, but involves complete modeling of human vocal system. VocalTractLab system as discussed in [2] is a good example of articulatory synthesizer. This system learns the articulation by analysis by synthesis method.
Formant Synthesizers were popular in 1980s but were abandoned in early 1990s. Klatt synthesizer [3] was one of the most widely used formant synthesizer. DECtalk and MITalk were the commercial versions of Klatt synthesizer. Early formant synthesizers produced robotic sounds due to limited processing cababilites of the processors. With the advent of modern programming laguages and powerful processors, development of formant synthesizers are again kindling interest in researches.

B. Why Formant Synthesis?

Concatenative synthesis appears as a lucrative solution to the problem of speech synthesis. Concatenative synthesis based TTS systems produces high quality output but the system on the whole becomes very huge and is not viable on smaller systems. Moreover, the actual problem of speech synthesis is traslated to a problem of searching the pre recorded sound units, matching prosodic features and sorting them in required order. A small change in the database can lead to degradation of quality of output. Also, they require huge amount of processing to change the prosodic features. Concatenative speech synthesizers do not throw any light on actual speech production mechanism as they do not need any type of modeling.
Articulatory synthesis is model based approach which requires indepth knowledge of human speech production system. Incorporating such vast knowledge in an implementation involves high development cost and time.
Formant synthesis approach is also a model based approach but is not as complex as articulatory synthesis. This method yields compact database and helps in understanding human vocal system from signal processing point of view. Availability of modern analysis and synthesis tools give the designers more liberty to experiment and to create better models to produce natural sounding utterances.

C. Klatt Type Formant Synthesizer

Simplified representation of Klatt synthesizer based on [4] is shown in Fig. 1. The synthesizer itself is based on source-filter model of speech production.

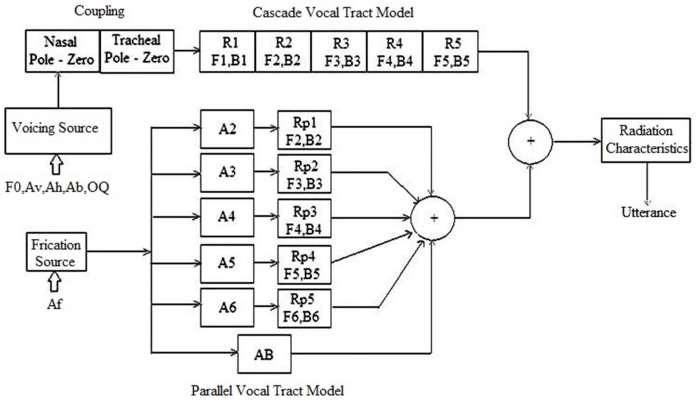

Figure 1. Klatt Type Formant Synthesizer

There are two excitation sources present. Voicing source produces synthetic rosenberg glottal pulse[5] having fundamental period F0 and is used for synthesizing vowels, diphthongs and semivowels. Frication source

produces noise and is used for synthesizing fricatives. Av and Af are the amplitude control parameter for voicing source and frication source respectively. Further the effect of breathiness and aspiration can be controlled using Ab, Ah and OQ controls as explained in [4]
The vocal tract is modeled using digital filters. Each digital filter is a second order all pole band pass filter and is called 'Resonator'. The bandwidth 'B' and centre frequency 'F' of each resonator depends on type of utterance being synthesized. Cascade vocal tract model is used for synthesizing voiced utterance and parallel vocal tract model is used for fricatives. The lip radiation is modeled as high pass filter.

D. Analysis of CV Coarticulation

The analysis of Consonant-Vowel (CV) coarticulation started as early as 1952. In a paper by A. Delattre, Pierre, Liberman and F. S. Cooper [6], it is identified that the transition of F2 formant contour during CV coarticulation from its consonant value to vowel value serves as cue for identifying stops and nasals. They also identified that the formant frequencies appears to originate from a single value for different vowel contexts for a consonant. Fig. 2. shows the spectrograms of plosive /ka/ formed by uttering consonant /k/ and vowel /a/ in isolation and by coarticulation.

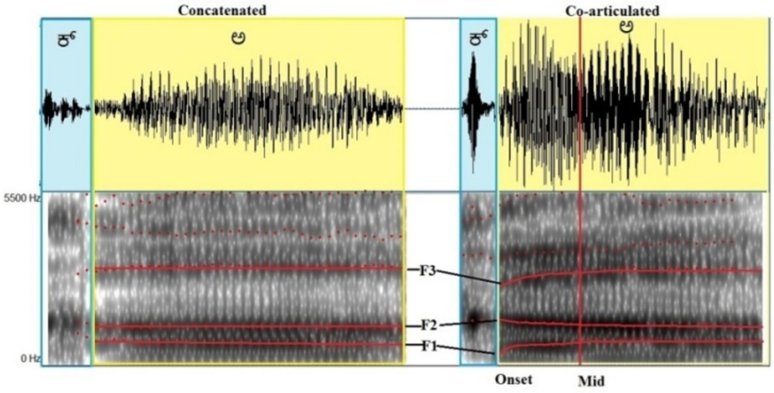

Figure 2. Formant contours for sound ಕ್(/k/) and ಅ(/a/ uttered in isolation and coarticulation

The formant frequencies of vowels are fixed and do not vary much with time if they are uttered in isolation. When the vowel appears in the context of a consonant, the formants begin with different values at their onset and reach their actual values in the middle of the utterance. This fact is evident in the spectrogram shown in Fig 2. The method of obtaining onset formant values in [7] and called this concept as 'Locus Theory'.
Equation (1) gives the relation between onset and mid values of i^{th} formant frequency of a vowel when it is coarticulated with a consonant.

$$Fi_{onset} = kF_{mid} + (1-k)Fi_{locus} \quad (1)$$

Equation (1) is essentially a regression line that fits the graph of F_{mid} Vs F_{onset} and the values of slope k and intercept
$(1\text{-}k)F_{locus}$ differs for various consonants. It is also evident that, the formant transition from their onset to mid frequencies follows the trend of (2)

$$Fi(t) = (Fi_{onset} - Fi_{mid})e^{-\alpha t} + Fi_{mid} \quad (2)$$

The formants of consonant also change by a small amount during transition. The values of i^{th} formant during consonant tranistion can be found using (3)

$$Fi(t) = (Fi_{cend} - Fi_{cstart})t + Fi_{cstart} \quad (3)$$

where Fi_{cstart} and Fi_{cend} are the starting and ending values of ith formant of a consonant during CV coarticulation.

II. Proposed Method

The goal is to generate Kannada isolated Consonant-Vowel coarticulations using Klatt like formant synthesizer. To generate CV utterances, it is just not enough to synthesize Consonant followed by Vowel and concatenate [8] and [9]. The consonant transition and vowel transition must also be implemented to get natural sounding utterances. To achieve this, process outlined in Fig. 3 was followed.

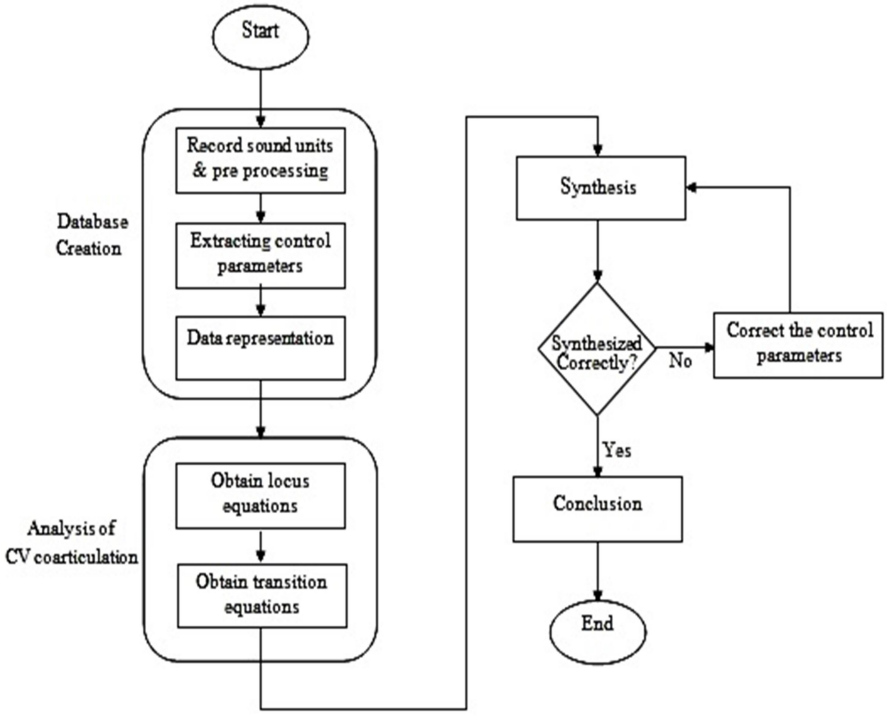

Figure 3. Workflow followed

A. Database Creation

Individual Kannada sound units such as vowels, diphthongs, consonants and consonant-vowels are recorded and preprocessed and stored. The control parameters are extracted as outlined in [10]. The control parameters such as pitch, formants, open quotient et cetera for isolated vowels, diphthongs and consonants are obtained from a male talker. To make the database compact, the contours are first normalized to a time duration of one second and then fitted to a polynomial. The coefficients of polynomial curve and actual time duration of the uttrences are stored.

Consonant Vowel utterences were recorded from three male talkers and two female talkers. For each consonant, onset and mid vowel frequencies of first three formants F1, F2 and F3 are measured with respect to 10 vowel context. Table I shows the same for consonant /k/ obtained from one talker. All the frequencies are measured in Hz.

TABLE I. SAMPLE ONSET AND MID VOWEL FORMANT FREQUENCIES FOR CONSONANT /K/

Vowel	F1 (onset)	F1 (mid vowel)	F2 (onset)	F2 (mid vowel)	F3 (onset)	F3 (mid vowel)
ಅ(/a/)	200	780	1404	1257	2594	2861
ಇ(/i/)	200	250	1179	2501	2768	3237
ಉ(/u/)	200	294	959	659	3884	2985
ಎ(/e/)	200	523	2443	2147	3299	2607
ಏ(/ee/)	200	404	2394	2268	3235	2808
ಒ(/o/)	200	514	1005	894	3028	3100

Along with onset and mid vowel formant frequencies, consonant transition duration, vowel transition duration are also tabulated. Fig. 4 shows an example of sound ಯ(/ya/). It can be observed that during consonant transition the formant frequencies of the consonant also change by a small amount. Hence it was decided to store starting and ending formant frequencies for consonant and its duration as well.

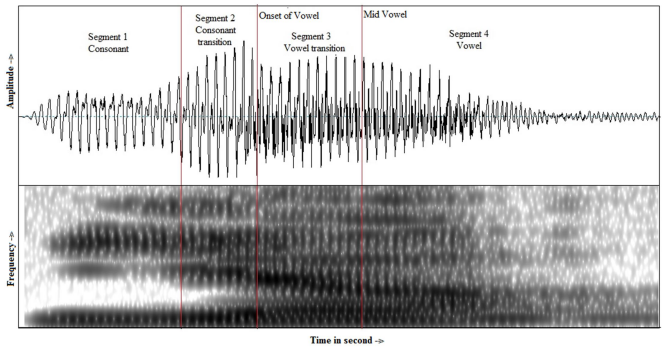

Figure 4. Waveform and spectrogram of ಯ(/ya/) showing different regions of CV coarticulation

B. Analysis of CV coarticulation

The locus equations are obtained by plotting onset frequencies versus mid vowel frequencies and fitting a linear regression to the points. It was also observed that only first three formants change during vowel formant transition and the higher formants were almost constant for the entire duration of vowel utterance. Fig. 5 shows regression line for sound /ka/.

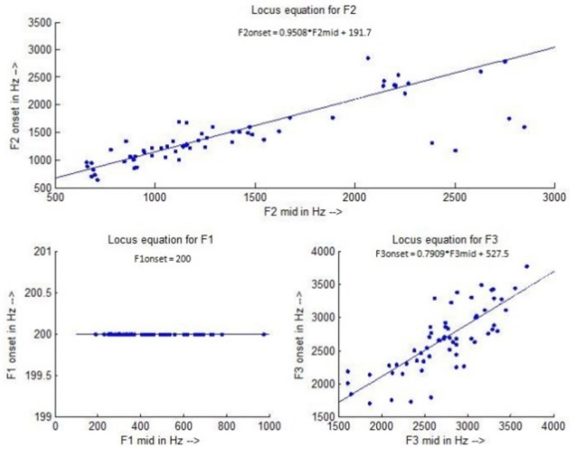

Figure 5. Regression lines for ಕ (/ka/)

Table II shows slope and intercept of (1) for some Kannada plosives. The average R^2 for F1 locus equation is 0.5102 and has a variance of 0.0479 and thus plays lesser role in coarticulation, for F2 locus equation average R^2 was found to be 0.7740 with variance 0.0257 and for F3, the average R^2 is 0.5637 with variance of 0.0297.

TABLE II. SLOPE AND INTERCEPT OF LOCUS EQUATION – PLOSIVES

Consonant	F1		F2		F3	
	Slope k	Intercept	Slope k	Intercept	Slope k	Intercept
ಕ್(/k/)	0	200	0.9508	191.7	0.7909	527.5
ಗ್(/g/)	0.3886	153.8	0.8577	367	0.7106	727.7
ಚ್(/ch/)	0	200	0.5256	1063	0.6619	1002
ಜ್(/j/)	0.2191	200.7	0.384	1327	0.5229	1488
ಪ್(/p/)	0.5918	124	0.7671	233.8	0.7808	494.6
ಬ್(/b/)	0.479	121.8	0.7546	251.9	0.666	856.2

It was observed that for soronants, the average R2 of F1 locus is 0.6838 with variance 0.0199 which is better compared to the case of plosives. Average R2 of F2 locus is 0.7549 having variance 0.0187 which is comparable to that of plosives. However F3 locus has lesser degree of importance in coarticulation compared to that of plosives with an average R2 of 0.4284 and variance 0.0447. Table III shows the details of soronant locus equations.

TABLE III. SLOPE AND INTERCEPT OO LOCUS EQUATION - SORONANTS

Consonant	F1		F2		F3	
	Slope m	Intercept	Slope m	Intercept	Slope m	Intercept
ಯ್(/y/)	0.7556	54.47	0.4766	1179	0.6185	54.47
ವ್(/v/)	0.6458	107.2	0.8294	96.35	0.8875	257.9
ರ್(/r/)	0.638	114.7	0.707	480.4	0.6508	805.6
ಲ್(/l/)	0.4796	161.7	0.5562	884.3	0.6602	977.6
ಳ್(/LL/)	0.4676	191.1	0.4372	1126	0.4881	1265

Table IV shows the slope and intercept of locus equations for some sililants, affricate and nasal.

TABLE IV. SLOPE AND INTERCEPT OF LOCUS EQUATION – SIBILANTS, AFFRICATE AND NASAL

Consonant	F1		F2		F3	
	Slope m	Intercept	Slope m	Intercept	Slope m	Intercept
ಸ್(/s/)	0.4784	159	0.494	902.6	0.882	373.2
ಶ್(/sh/)	0.4324	129	0.4493	1225	0.7176	931.9
ಹ್(/h/)	0.8311	50.87	0.971	58.16	1.027	-58.69
ಮ್(/m/)	0.6841	41.18	0.8193	277.1	0.3438	1870

C. Synthesis Strategy

Most of the Klatt synthesizers use fixed frame size anywhere between 5ms to 25ms and update the parameters once per frame. Speech is generally assumed to be stationary in those frames. If frame size is very large, then parameters change drastically when updated yielding robotic speech. To avoid that, control parameters are updated once every pitch period as outlined in [10]. Following algorithm is used to synthesize isolated vowels, diphthongs and CV utterances

Step 1: Set time t = 0 and set synthesis duration d.
Step 2: Evaluate pitch contour say $P(\tau)$ and other parameter contours at $\tau = t/d$. Calculate pitch period say $\Delta = 1/P(\tau)$.
Step 3: Apply all the parameters to the synthesizer and synthesizer for Δ amount of time.
Step 4: Set time $t = t + \Delta$
Step 5: Go back to step 2 if time t is less than synthesis duration d.
Step 6: Stop

The isolated vowels and diphthongs can be synthesized in a single segment using above algorithm. However CV utterances are synthesized in multiple segments. The control parameters for synthesizing Segments 1 and 4 as shown in Fig. 4 are directly available in database. For consonant formant transition, instead of regular F1, F2 and F3 contours, Equation (3) is evaluated by extracting Fi_{cstart} and $Fi_{cend,}$ i = 1,2,3 for given consonant from database. For vowel formant transition, (2) is evaluated. Fi_{mid} is evaluated from the regular formant contour at indicated value of time. Once Fi_{mid} is obtained, Fi_{onset} can be calculated from the locus equation of the consonant.

III. RESULTS

Ten isolated vowels and two isolated diphthongs were synthesized using above described method. The synthesis of vowels and diphthongs do not need any type of transition equations to be evaluated. 15

consonants with 10 vowels and 2 diphthongs contexts totalling 180 CV utterances were also synthesized. Fig. 6 shows spectrongram of synthesized sound ಗ(/ga/).

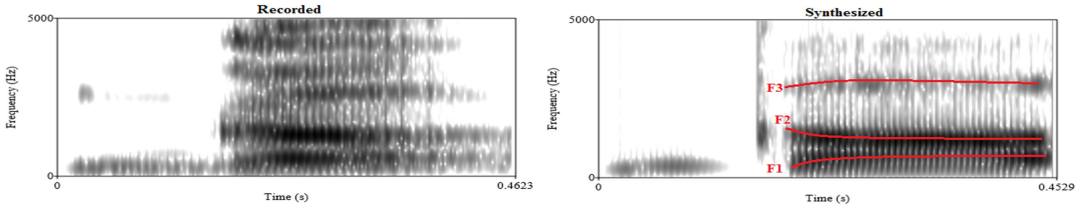

Figure 6. Spectrograms of recorded and synthesized waveform of ಗ(/ga/)

Unlike vowels and diphthongs, CV utterance generation requires, generation of consonant C in the first segment, if there are any consonant formant transition then it is generated in second segment. Based on the vowel context, onset frequencies are calculated using locus equations of consonant C. Once onset frequencies are obtained, vowel formant transition contour can be formed using transition equations. An example of such CV utterance is shown in Fig. 6. The first three formant contours are marked and shows clearly all the segments.

A perception test was conducted to evaluate the intelligibility of utterances, three listeners were asked to identify the isolated utterances. An utterance was declared as correctly identified if all the three listeners identified it. Even if one of the three listeners failed to identify an utterance, it was declared as wrongly. Table V shows the result of perception test.

TABLE V. RESULT OF PERCEPTION TEST

Category	Total Utterances	Correct Identification	Wrong Identification	Percentage Correctly Identified
Vowels	10	10	0	100
Diphthongs	2	2	0	100
Plosives	72	68	4	94.44
Soronants	60	56	4	87.5
Sibilants & Affricate	36	34	2	94.44
Nasal	12	6	6	50

A total of 192 utterances were synthesized, out of which 176 (91.67%) were correctly synthesized. Out of 16 wrongly synthesized utterances, 4 were in the context of back vowels, 10 in the context of front vowels and 2 in the context of central vowels. The analysis of CV utterances in this project leads to following conclusions.

1. Only first 3 formants play an important role in CV co-articulation. F4 and F5 formant transitions are barely noticeable
2. F2 formant transition has relatively low tolerance of error compared to F1 and F3 transitions.
3. F3 onset frequencies have less span and tends to concentrate around a common point.
4. Locus equations for Kannada language are different than that of other languages.

IV. CONCLUSION AND FUTURE SCOPE

This work is one of the stepping stones towards designing a functional Kannada Formant Synthesizer based Text-To-Speech conversion system. All the control parameters required for generating basic sound units of Kannada language are evaluated and tabulated during the course of this project. The locus equations and formant transition equations found for CV utterances are helpful in analyzing and implementing intonation using formant synthesizers.

Some of the areas which are not considered in this work due to various constraints. Each of those areas are a necessity for a fully functional TTS system.

1. Analysis of aspirated plosives and remaining nasal Consonant-Vowel utterances was not taken up in this work and is left open for further investigation.
2. Consonant-Consonant-Vowels coarticulations (ಒತ್ತಕ್ಷರ) also plays a dominant role in Kannada language.

3. Vowel to Consonant (VC) transition also plays an important role in continuous speech generation which needs a thorough investigation.

REFERENCES

[1]. Sarathy, K. Partha, and A. G. Ramakrishnan. "A research bed for unit selection based text to speech synthesis." *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*. IEEE, 2008.

[2]. Prom-on, Santitham, Peter Birkholz, and Yi Xu. "Training an articulatory synthesizer with continuous acoustic data." *INTERSPEECH*. 2013.

[3]. D. H. Klatt, "Software for a cascade/paralell formant synthesizer," *J. Acoust. Soc. Am.*, vol. 67, no. 3, pp. 971–995, 1980.

[4]. D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers.," *J. Acoust. Soc. Am.*, vol. 87, no. 2, pp. 820–857, 1990.

[5]. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels.," *J. Acoust. Soc. Am.*, vol. 49, no. 2B, pp. 583–590, 1971.

[6]. Delattre, Pierre, Liberman and F. S. Cooper, "Acoustic Loci and Transitional Cues for Consonants," J. Acoust. Soc. Am., vol. 7, no. 4, pp. 769–773, 1955.

[7]. D. H. Klatt, "Review of text-to-speech conversion for English.," J. Acoust. Soc. Am., vol. 82, no. 3, pp. 737–793, 1987.

[8]. X. A. Furtado and A. Sen, "Synthesis of unlimited speech in Indian languages using formant-based rules," *Sadhana*, vol. 21, no. 3, pp. 345–362, 1996.

[9]. Kang, Shinae, Keith Johnson, and Gregory Finley. "Effects of native language on compensation for coarticulation." *Speech Communication* 77 (2016): 84-100.

[10]. V. D'Souza and Dr. D.J Ravi. "Modified Synthesis Strategy for Vowels and Semivowels(Klatt Synthesizer)," International journal of Electronics and Communication Engineering &Technology (IJECET).Vol. 5,no 8,pp. 61-70, 2014