International Journal of Electronics Engineering, 2(1), 2010, pp. 75-81

Effect of Consonant Duration Modifications on Speech Perception in Noise-II

NH Shobha[1], TG Thomas[2] & K Subbarao[3]

[1]*Research Scholar, Dept. of ECE, Osmania University, Hyderabad, India-500007*
[2]*Professor., Dept. of Electrical Engineering. BITS-Pilani, Dubai, UAE*
[3]*Professor., Dept. of ECE, Osmania University, Hyderabad, India-500007*

Abstract: The paper is an extension of our previous paper which addressed the efficacy of temporal modifications to the clear speech advantage. A case for synthetic clear speech in the context of hearing impairment was developed on Fricative-vowel syllables. Stimuli were subjected to consonant-duration lengthening, where fricative noise duration and formant transition segments were time-expanded independently by 50-100% of their original duration. The speech perception in noise (SPIN) tests were quantified in terms of information transmission analysis measures, in the presence of white noise-masker at three noise levels, 0 dB, +12 dB, and +6 dB. The findings reported that lengthening formant transition duration by 50% was found to improve speech intelligibility in simulated high level sensorineural hearing loss while fricative noise duration had no positive benefit.

Keywords: Consonant-duration Modification, Information Transmission Analysis, White-noise Masker, Speech Intelligibility

INTRODUCTION

People suffering from hearing loss are often said to have greatest difficulty in identifying short speech sounds such as plosive consonants and fricative consonants. It is also widely known that consonants are less intelligible than vowels, as they are weaker in strength and shorter in duration compared to vowels [1]. The consonants especially within the same class are often difficult to differentiate and are more vulnerable to signal degradations hence, it is desirable to strengthen the available acoustic cues to make consonant contrasts more distinct and potentially more robust to subsequent noise degradations. Our previous work [2] has shown the effects of CD modifications (CDM) on speech intelligibility using Plosive-vowel syllables; and the present work extends CDM on to Fricative-vowel syllables.

The plosives and fricative consonant sounds are produced by a constriction in the vocal tract. Fricatives are those consonants produced when the turbulent air-flow occurs at a point of constriction in the vocal tract. Fricative consonants are characterized by a turbulent noise, and may consist of the noise alone or may consist of the noise together with vocal cord vibration. Fricatives and plosives bursts are both characterized by high frequency random noise, which occurs on the opening of oral cavity. Plosives are characterized by highly transient cues, release burst very brief; whilst the noise spectrum of a fricative is quiet a great deal longer and rises to its target amplitude more gradually than a plosive does.

*Corresponding Author: [1]Shobhanh7@gmail.com, [2]tgth@hotmail.com, [3]kakarlasubbarao@yahoo.com

Much research in the past has predominantly focused on the perception of plosive consonants; in contrast, clearly produced fricative consonants have not been the subject of more observation. Moreover, it is uncertain whether the classification metrics proposed for stop consonants can be successfully applied to fricatives. The current work focused on lengthening consonant duration of fricative consonants. Lengthening speech sounds is said to produce slow speech rate, thus provides extra processing time for the hearing impaired subjects.

Nonsense syllables involving fricative consonants {/f/, /θ/, /s/, /v/, /ð/, /z/} in CV context with cardinal vowels /a, i, u/ were used as test stimuli. The main acoustic cues that have been reported to affect perception of fricatives (for normal hearing listeners) include - Noise Duration (ND) and amplitude, as well as adjacent Formant Transition Duration (FTD) [3]. The 'Noise Duration' is said to extend from fricative onset time to fricative offset time, the time interval between the onset of the following vowel and the instance when a formant frequency reaches its steady-state value is called the 'Formant Transition.

In the current investigation, natural syllables were recorded, subjected to resynthesis, processed for consonant duration lengthening in two independent schemes, (i) Noise Duration Modification (NDM), and (ii) Formant Transition Duration Modification (FTDM). The consonant lengthening protocol employed was PSOLA (pitch-synchronous overlap and add) [4], where the original pitch is being preserved during the processing [5]. The PSOLA analysis-modification-synthesis method belongs to the general class

of STFT (Short-Time Fourier Transform) analysis-synthesis method.

1. SPEECH IN NOISE TASK

1.1. Speech Material

Nonsense syllables with consonant-vowel /CV/ structure were chosen for investigation. The idea behind this non sense syllable test (NST) was to maximize the contribution of acoustic factors, and minimize the impact of adjacent vowels. The test material consisted of fricative consonants - {/f/, /θ, /s/, /v/, /ð/, /z/} with primary cardinal vowels {/a/, /i/, /u/} forming voiceless /CV/ subset, {/fa/, /fi/, /fu/, /θa/, /θi/, /θu, /sa/, /si/, /su/} and voiced /CV/ subset, {/va/, /vi/, /vu/, /θa/, /θi/, /θu/, /za/, /zi/, /zu/}.

1.2. Speech Signal Processing

The signal processing was accomplished in four different stages as explained below. In the first stage, we recorded the natural speech tokens and subjected them to resynthesis. The natural stimuli were recorded in a quiet room, sampled at 44.1KHz, using a PRAAT monosound recorder. The best utterance out of 20 utterances of the first author (middle aged, female) was selected based on the proper phonetic clarity.

The speech tokens were subjected to resynthesis using the procedure of LPC (linear prediction) analysis-synthesis as provided in PRAAT [6]. The idea behind the resynthesis was two-fold; firstly, the synthetic copy renders efficient and independent manipulation of the spectral, temporal and intensity characteristics; secondly, synthetic speech is as similar as possible to a human utterance. For implementing linear prediction with Praat, we have to implement this band-limiting by resampling the original signal to 11 KHz for female, 10 KHz for male, or 20 KHz for a young child. In the current investigation, we performed resampling at 11 KHz, extracted the filter and the source from the resampled sound using linear-prediction analysis. The analysis procedure adopted 10 linear-prediction coefficients (yields at most 5 formant-bandwidth pairs) in each time frame of 5 or 10 ms, which is suited for capturing changes in the speech signal. Next, using the extracted source and filter, the speech sound was regenerated based on LPC synthesis. This procedure gave back the resynthesized version with the original quality except that the windowing caused few ms at the beginning and the end of the signal to be set to zero Finally, these tokens were normalized to 70 dB IL (referred as baseline syllables) to avoid the signal clipping in subsequent processing stages.

In the second stage of processing, consonant segment durations such as fricative ND and FTD measurements were measured by visual inspection of the time waveforms and wideband spectrograms using the PRAAT software. Noise Duration is referred to the high frequency noise, measured as the difference between the fricative offset time and fricative onset time. The fricative onset time is the point at which high frequency energy appeared on spectrogram and/ or point at which the number of zero crossings rapidly increased, while the fricative offset time is the intensity minimum immediately preceding the onset of vowel periodicity, for voiceless fricatives and the earliest pitch period exhibiting a change in waveform from that seen throughout the initial frication, zero crossing of the preceding pitch period was designated as fricative offset, for voiced fricatives [7].

The Formant Transition Durations are measured by simultaneous consultation of time domain waveform, spectrogram, linear-predictive coding (LPC) spectra, and short-time fast-fourier transform (ST-FFT) spectra [8] The LPC spectrum was constituted for a prediction order of 10 (at least twice as the number of spectral peaks that we want to detect), analysis window of 12.5 ms and 5 ms step, +6dB/ octave filtering above 50 Hz. The three formants were originally located by examining the LPC spectra, FFT spectra, and spectrogram. The steady-state point of the vowel was centered at 100 ms after the onset. Formant analysis was performed for the detection of formant transition duration. After proper settings, formant contour was extracted and the formant values were written to a text file. Utilizing this data, the duration of the transitions and their onset and offset points were determined, and we then applied a time warp to all formants over the determined duration of the transition. The acoustic segmentations and measurements were done using PRAAT software.

In the third stage of processing, the extracted acoustic segments were subjected to duration modification or time-stretching. This stage of processing employed a time-stretching algorithm referred as Pitch-Synchronous Overlap and Add (PSOLA). Based on the modification strategies, consonant duration modifications took three different schemes; (i) Noise duration modification – NDM, and (ii) Formant Transition Duration modification – FTDM. The PSOLA analysis-modification-synthesis method belongs to the general class of STFT (short-time Fourier Transform) analysis-synthesis method. In the PRAAT object window, PSOLA can be found as sound>Convert>Lengthen (PSOLA). Here, the term 'factor' decides the factor for lengthening or shortening; by choosing factor value >1 or <1, the resulting sound could be longer or shorter than the original segment, but a factor value larger than 3 will not work. We selected a minimum pitch of 75Hz and a maximum pitch of 600Hz, while a 'factor' of 1.5 for 50% lengthening(compared to original duration) and a 'factor' of 2 for 100% lengthening(compared to original duration). Finally, the lengthened segment was blended back to its original location to result in time stretched version. We thus obtained three modifications for each stimulus, one without modification (0%) and other two with modifications (50% and 100%) under all three schemes (NDM/FTDM).

The fourth stage of processing was designed to simulate hearing impairment, by reducing the acoustic dynamic range. The masking noise responsible for the threshold elevation is believed to be predominantly of cochlear origin [9]. As reported in literature, the reduction in the hearing threshold can be approximately simulated by addition of white noise [10, 11]. Some researchers have employed multi-talker babble instead of white noise [11, 12, and 13]. However, due to its non-stationary nature, the effective masking it may provide during stimulus presentation is unpredictable. Hence, we decided to use white noise masker to model the hearing loss to a good approximation.

The processed tokens from the previous stage were additively mixed with the synthesized noise at three noise conditions, i.e., no-masking noise, +12 dB and +6dB SNRs. The noise free (natural) tokens were considered as no-masking noise tokens. The SNR refers to the ratio of the average power in CV token to the average power of the noise token in decibels. For deriving +12dB and +6dB SNR-tokens, the average power level of the speech token was fixed while that of the noise was adjusted. PRAAT scripts were run for synthesizing the white noise and for the process of mixing [14]. The Chris-Darwin [14] algorithm which performed additive mixing summed up the sounds by point-to-point values, preserving real time across the time domains. Finally, after four stages of processing stimuli corpus holds 324 test tokens spanning across 18 syllables, 2 duration lengthening schemes, 3 versions of lengthening per scheme and 3 SNRs per version.

1.3. Speech Perception in Noise

Two female and two males in the age group of 16-45 years with normal hearing, participated in the listening experiments. None of the subjects were experienced with perceptual experiments; subjects went through a stimuli familiarization sessions before the experiment started.

The perception tests were automated using a MATLAB code with graphic user interface. Stimuli were presented using a computerized testing procedure at the most comfortable listening level of 75 to 85 dB SPL for the listeners. The test procedure used a similar protocol for all three experiments. The Experiment under each individual scheme (NDM/ FTDM*)* worked on a total of 162 tokens categorized under 9 or (3*3) listening conditions. These included the original and processed stimuli with 3 levels of CD lengthening (0%, 50% and 100%) and 3 noise levels (no-noise, +12dB, +6dB). Under every listening condition, subjects were played tokens with ten randomized replications of each token; they were prompted to choose from the set of choices displayed on the computer screen. Results were cast into three groups of six by six confusion matrices (CM) per run.

1.4. Speech Intelligibility Measures

Speech discrimination test results were summarized as the percentage of correct responses for many experimental runs. We adopted the Information Transmission analysis approach [15, 16, and 17], which provides a measure of covariance between stimuli and responses, and takes into account the pattern of errors and the score in a probabilistic manner. The covariance measure of intelligibility can be applied to the sub matrices derived from the original matrix by grouping the stimuli in accordance with certain desired features [16, and 17]. The information measures of the input stimulus X and output response Y are defined in terms of the Mean Logarithmic Probability - MLP, given by,

$$I(X;Y) = -\sum_i \sum_j P(x_i, y_j) \log_2 \left(\frac{p(x_i)\, p(y_j)}{p(x_i, y_j)} \right) \text{bits} \quad (1)$$

The Relative Information Transmission (RIT) from *X* to *Y* is given by,

$$I_{tr}(X;Y) = \frac{I(X;Y)}{I_s(X)} \quad (2)$$

Where, *Is*(*x*) is the information measure of the input-stimulus in terms of MLP.

2. EFFECTS OF CONSONANT DURATION MODIFICATIONS ON SPEECH INTELLIGIBILITY

The confusion matrices obtained were analyzed and quantified with perceptual (information transmission analysis) and statistical (two-tailed t-test) measures. The perceptual scores were obtained by averaging the scores for individual subject across three vowel contexts /a, i, u/; the last rows in the table indicates their means and standard deviations. The statistical tables reported the mean percent-correct recognition data, standard deviations (SD), probability value (p) and the corresponding statistical significance value corresponding to the perception test. The processing factor examined the intelligibility benefit between the unprocessed speech and the processed speech, a benefit was treated significant at 0.05 levels; $p<=0.01$ was accepted as indicative of 'high significance' and $0.05<p<0.01$ as 'moderate significance', and $0.1<p<0.05$ as 'minimal significance'.

2.1. NDM Paradigm

Tables 1(a) and 1(b) represents the perceptual analysis and statistical analysis scores respectively. For no-noise presentations, voiceless and voiced fricatives reported negative benefits for a majority conditions. Similarly, in the presence of +12dB and +6dB noise, the effect was reported to be detrimental due to majority negative benefits. The statistical analysis (table 1(b)) presents the significance status of the perceptual measures (table 1(a)). The analysis has reported no significant benefit under for all three SNR presentations, and three ND modifications.

2.2. FTDM Paradigm

Tables 2(a) and 2(b) represent the perceptual analysis and statistical analysis score pattern for FTDM scheme. The statistical test (table 2(b)) reports the significance status of the perceptual measures (table 2(a)). No-noise presentations and +12 dB SNR presentations, did not report any significant benefit either for 50% or 100% FTDM's. But +6dB presentations, reported highly significant benefit ($p<=0.01$) corresponding to 50%FTDM (voiced fricatives) and minimally significant benefit ($0.1<p<0.05$) corresponding to 50%FTDM (voiceless fricatives); while 100%FTDM did not report significant benefit for voiced/voiceless fricatives.

3. SUMMARY AND CONCLUSIONS

The findings suggested that of the two acoustic segment modifications considered here, increase in FTD have reported positive intelligibility benefits at high-level masking noise (+6dB); however, ND did not appear to be suitable for consonant duration modification. Therefore, in the presence of noise or for the benefit of the hearing-impaired listeners, FTD is found to be a more dominant cue for lengthening consonant duration of fricative consonants.

In total, based on present work and previous work [2], we can conclude that consonant duration lengthening study enjoyed a significant intelligibility benefit in noise for (i) Burst Duration lengthening by 50% of its original duration (for plosives), (ii) FTD lengthening by 100% of its original duration (for plosives), and (iii) FTD lengthening by 50% of its original duration (for fricatives). The findings suggested that efforts to emphasize potentially weak consonant should be beneficial in surmounting some of the speech recognition difficulties of hearing impaired listeners.

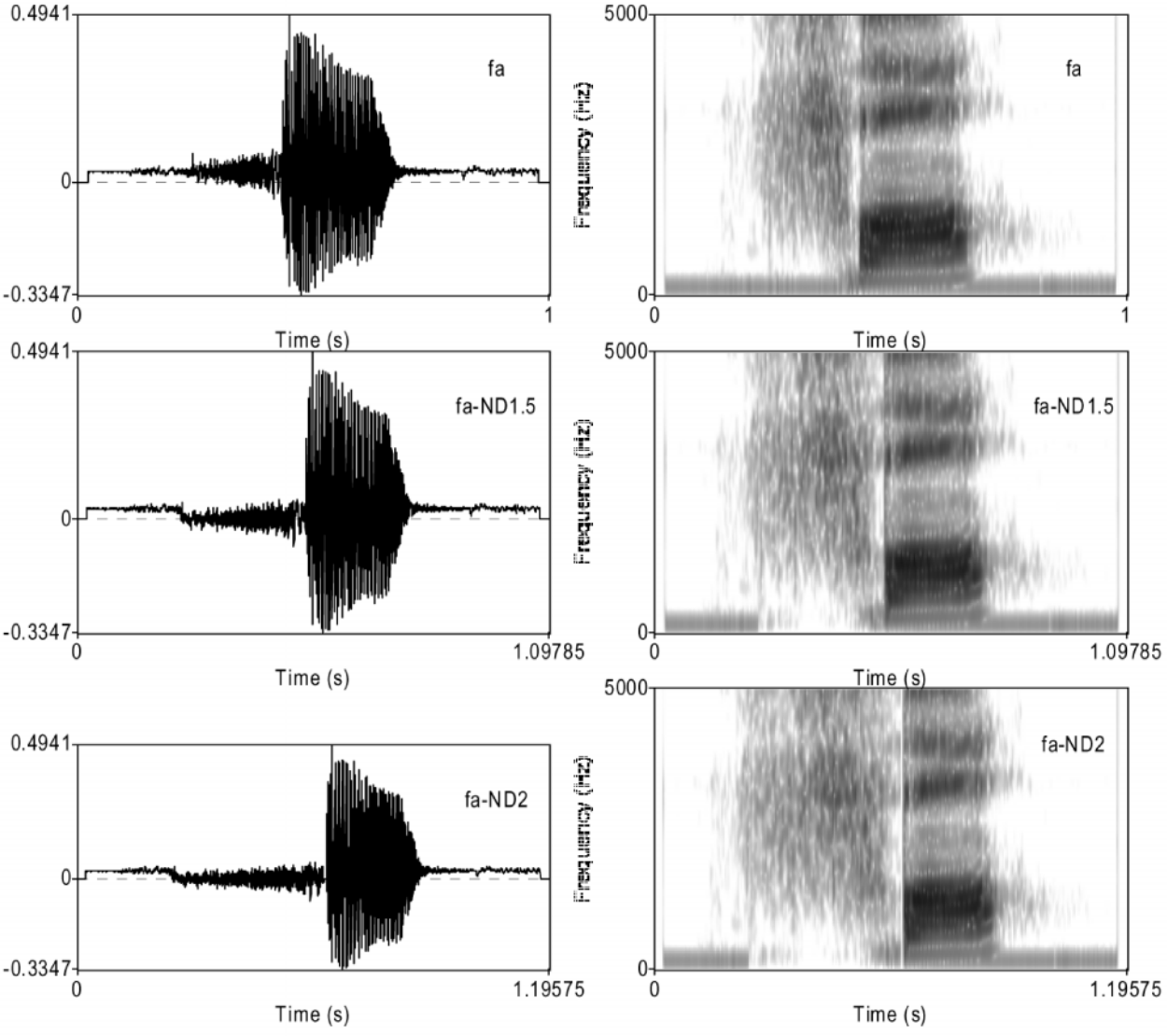

Figure 1: NDM Paradigm: Temporal Waveforms and Spectrograms for /fa/ Syllable in the Normal and Two Modified Conditions

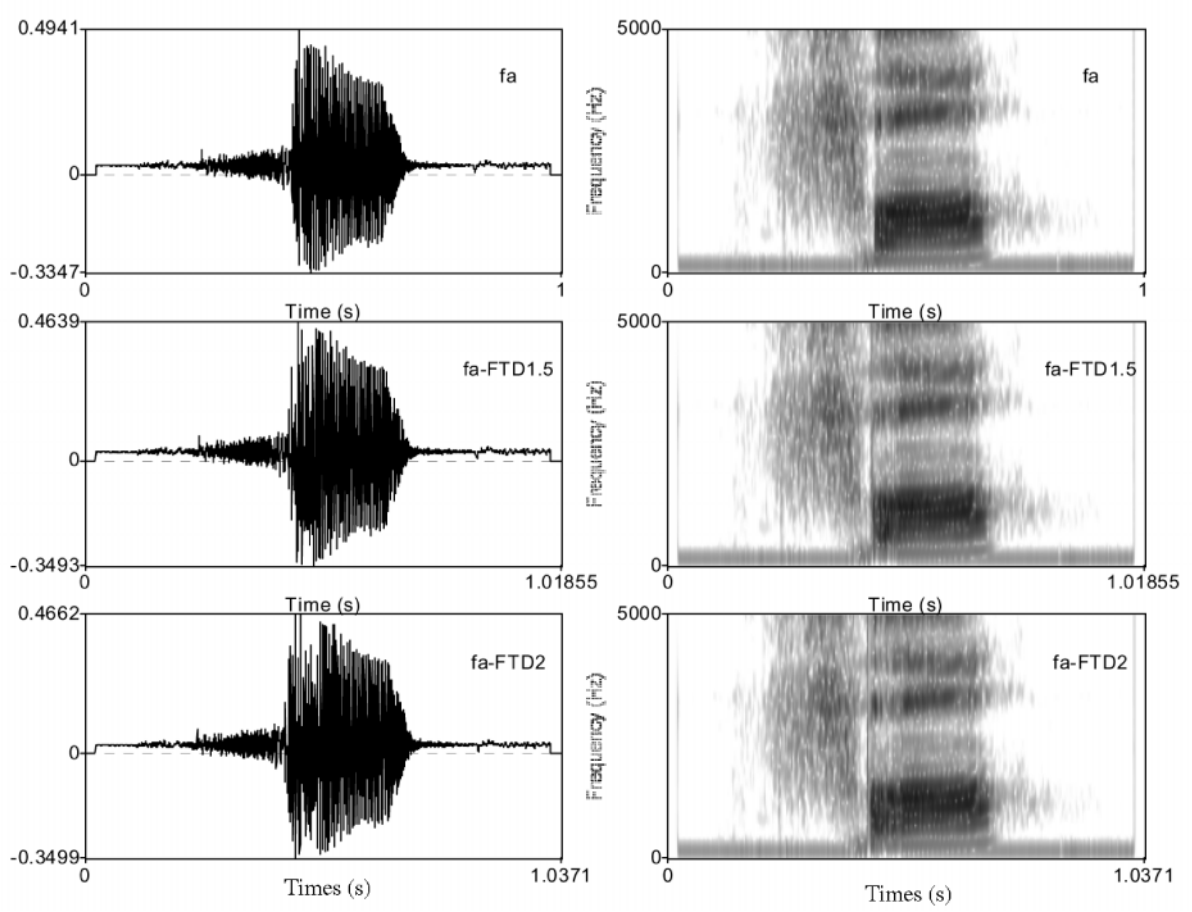

Figure 2: FTDM Paradigm : Temporal Waveforms and Spectrograms for /fa/ Syllable in Normal and Two Modified Conditions

Table 1(a)
NDM Scheme- Perceptual Analysis

Test stimuli	*Listener*	*Relative information transmitted (%) (%)*								
		No masking noise			*SNR = 12 dB*			*SNR = 6 dB*		
		NDM (%)			*NDM (%)*			*NDM (%)*		
		0	50	100	0	50	100	0	50	100
	L1	100	100	100	83	100	97	83	88	81
Voiceless	L2	100	100	97	89	89	89	84	88	81
Fricative-	L3	93	100	100	87	83	92	76	79	88
vowels	L4	100	100	97	100	97	61	78	86	78
	AVG	98	100	98	90	92	85	80	81	83
	SD	3	0	2	7	7	16	4	8	5
	L1	100	97	100	100	100	100	100	100	91
Voiced	L2	100	100	100	100	100	100	100	93	95
Fricative-	L3	100	97	100	100	100	100	100	88	100
vowels	L4	100	97	97	97	97	93	93	97	93
	AVG	100	97	99	99	99	98	98	95	95
	SD	0	2	2	2	2	3	3	5	4

Table 1(b)
NDM Scheme-Statistical Analysis

Test Stimuli	SNR (dB)	NDM (%)	Mean	SD	Two Tailed t Test of Difference		
					t	p	Result
Voiceless Fricative-vowels	No-noise	0	98	3			
		50	100	0	NaN	NaN	ND
		100	98	2	0	1	NS
	12	0	90	7			
		50	92	7	0.404	0.7002	NS
		100	85	16	-0.573	0.5877	NS
	6	0	80	4			
		50	81	8	0.224	0.8305	NS
		100	83	5	0.937	0.3849	NS
Voiced Fricative-vowels	No-noise	0	100	0			
		50	97	2	NaN	NaN	ND
		100	99	2	NaN	NaN	ND
	12	0	99	2			
		50	99	2	0	1	NS
		100	98	3	-0.555	0.5992	NS
	6	0	98	3			
		50	95	5	-1.029	0.3432	NS
		100	95	4	-1.2	0.2754	NS

Table 2(a)
FTDM Scheme-Perceptual Analysis

Test stimuli	Listener	Relative information transmitted (%) (%)								
		No masking noise			SNR=12 dB			SNR=6 dB		
		FTDM (%)			FTDM (%)			FTDM (%)		
		0	50	100	0	50	100	0	50	100
Voiceless Fricative-vowels	L1	100	97	97	83	97	100	83	73	75
	L2	100	100	100	89	94	89	84	78	75
	L3	93	100	100	87	97	86	76	77	68
	L4	100	100	100	100	100	97	78	72	81
	AVG	98	99	99	90	97	93	80	75	75
	SD	3	2	2	7	3	7	4	3	5
Voiced Fricative-vowels	L1	100	100	97	100	93	93	100	83	100
	L2	100	100	100	100	97	97	100	88	88
	L3	100	97	100	100	100	100	100	81	100
	L4	100	97	100	97	100	97	93	84	93
	AVG	100	98	99	99	97	97	98	84	95
	SD	0	2	2	2	3	3	3	3	6

Table 2(b)
FTDM Scheme-Statistical Analysis

Test Stimuli	SNR (dB)	FTDM (%)	Mean	SD	Two Tailed t Test of Difference		
					t	p	Result
Voiceless Fricative-vowels	No-noise	0	98	3			
		50	99	2	0.555	0.5992	NS
		100	99	2	0.555	0.5992	NS
	12	0	90	7			
		50	97	3	1.838	0.1157	NS
		100	93	7	0.606	0.5667	NS
	6	0	80	4			
		50	75	3	-2	0.0924	NS
		100	75	5	-1.562	0.1694	NS
Voiced Fricative-vowels	No-noise	0	100	0			
		50	98	2	NaN	NaN	ND
		100	99	2	NaN	NaN	ND
	12	0	99	2			
		50	97	3	-1.109	0.3097	NS
		100	97	3	-1.109	0.3097	NS
	6	0	98	3			
		50	84	3	-6.6	0.0006	S
		100	95	6	-0.894	0.4055	NS

REFERENCES

[1] P Ladefoged, "Vowels and Consonants: An Introduction to the Sounds of Language", Oxford: Blackwell Publishing, 2001.

[2] Shobha, N.H., Thomas, T.G., and Subbarao, K., "Effect of Consonant Duration Modifications on Speech Perception in Noise", *Int. J. Electr. Engg.*, **1(2)**, pp 179-184, Dec 2009.

[3] Van Heuven, V.J., "Reversal of the Rise-time Cue in the Affricative-fricative Contrast: An Experiment on the Silence of Sound", *The Psychophysics of Speech Perception*, Martinus Nijhoff, Dordrecht, pp 181-187, 1987.

[4] Moulines, E. and Charpentier, F., "Pitch-synchronous Waveform Processing Technique for Text-to-speech Synthesis using Diphones", *Speech Communication*, **9(5)**, pp 453-467, 1990.

[5] Liu, S. and Fan-Gang Zeng. : "Temporal Properties in Clear Speech", *J. Acoust. Soc.Am.*, **120(1)**, pp 424-432, 2006.

[6] Boersma, P. and Weenik, D.: "Praat: Doing Phonetics by Computer" Version 4.4.22, 2005.

[7] Jongman, A., Spence, M., Wang, Y., Kim, B., and Schenck, D., "Perceptual Properties of English Fricatives", *J. Accost. Soc. Am.*, **105**, 1401(A), 1999.

[8] Jiang, J. Chen, M. and Alwan, A.: "On the Perception of Voicing in Syllable-initial Plosives in Noise", *J. Acoust. Soc. Am.*, **119(2)**, pp 1092-1105, 2005.

[9] Oxenham, A.J. and Moore, B.C.J.: "Modeling the Effects of Peripheral Nonlinearity in Listeners with Normal and Impaired Hearing", In W. Jesteadt (Ed.), "Modeling Sensorineural Hearing Loss"(Mahwah, NJ: Lawrence Earlbaum Associates), pp 273-288, 1987.

[10] Sussman, J.E.: "Perception of Formant Transition Cues to Place of Articulation in Children with Language Impairments", *J.Speech Lang. Hear.Res.*, **30**, pp 1286-1299. 1993.

[11] Flecher, H.: "The Perception of Sounds by Deafened Persons", *J. Acoust. Soc. Am.*, **24**, pp 490-497, 1952.

[12] S A Simpson, & M Cooke: "Consonant Identification in N-talker Babble in a Non Monotonic Function of N", *J. Acoust. Soc. Am.*, **118(5)**, pp 2775-2778, 2005.

[13] H D Lewis, V A Benignus, K E Muller, C M Malott, "Babble and Random Noise Masking of Speech in High and Low Content Cue Identifications", *J. speech and Hearing Res*, **31**, 108-114, 1988.

[14] <http://www.lifesci.Sussex.ac.uk/home/Chris_Darwin/praatscripts/ >

[15] Nittrouer, S.: "Do Temporal Processing Deficits Cause Phonological Processing Problems?", *J. Speech Lang*, Hear. Res., **42**, pp 925-942. 1999.

[16] Bilger, R.C. and Wang, M.D.: "Consonant Confusions in Patients with Sensorineural Hearing Loss", *J. Speech and Hearing Res*., **19**, pp 718-748, 1976.

[17] Miller, G.A. and Nicely, P.E.: "An Analysis of Perceptual Confusions among Some English Consonants", *J. Acoust. Soc. Am.*, **27(2)**, pp 338-352, 1955.