The effect of speaking rate and vowel context on the perception of consonants in babble noise

Anirudh Raju

Department of Electrical Engineering , University of California, Los Angeles, California, USA

anirudh90@ucla.edu

Abstract

In this paper, we study human perception of consonants in the presence of additive babble noise at two speaking rates. In addition, we work on a model that attempts to replicate these human results through a phoneme recognition model. Consonant-Vowel-Consonant (CVC) stimuli comprising of a set of 13 consonants and 3 vowels (/a/, /i/, /u/) were recorded in a sound proof booth by two talkers at two different speaking rates (fast and slow). Noisy stimuli were generated by adding babble noise at different levels to the quiet recordings. These stimuli were used to conduct perceptual experiments in which 52 listeners were asked to listen and repeat back the CVC phrases presented in babble noise under 3 SNR conditions and both speaking rates. The data was transcribed by two trained linguists. The results were analyzed by SNR, vowel, and speaking rate. Rate did not have an effect on the perception of consonants in quiet conditions. With the exception of /CuC/ stimuli, speaking rate had a pronounced effect, with slow speech being more intelligible than fast speech in the presence of noise. /CaC/ stimuli were, on average, more robust than other stimuli in all conditions. In addition, syllable initial consonants were better identified than final consonants, especially in noise. The effect of rate was more pronounced in voiced syllable final consonants. The stimuli collected was run through a phoneme recognition model based on GMMs. The machine effectively sees ”noisy” data even though we present it with clean stimuli due to mismatched train and test conditions (train on TIMIT, test on sound booth stimuli). The machine consonant recognition accuracies and confusions are similar to that of humans for 0dB stimuli. The machine also mimicks the effect of rate on humans for the /i/ and /u/ context while it does not for the /a/ context..

Index Terms: Rate, Perception, CVCs, Confusion Matrix

1. Introduction

In the presence of a noise masker, speech sounds are often confused with others by both human listeners and machines. Hence, a big challenge for automatic speech recognition (ASR) systems and human listeners is to perceive speech in the presence of background noise. It is useful to study the way humans perceive speech in noise in order to enable us to build more noise robusts systems for the same. Nonsense syllables of the form of Consonant-Vowel-Consonant (CVC) or Consonant-Vowel(CV) or Vowel-Consonant(VC) are typically used in perceptual experiments to study the way humans confuse consonants and vowels in the presence of noise at different SNRs. The classic study [1] performed experiments with CV syllables (16 consonants, 1 vowel, in white noise). It showed that some consonants are identified more easily than others at the same SNR. The percentage accuracy of a consonant was used as a metric of consonant identification. A more detailed analysis on the perceptual confusion patterns were studied in [2].

Following the classic study [1], there have been several studies on consonant confusions in the presence of noise. A phonetically balanced set of nonsense English syllables containing ten initial consonants, ten vowels and ten final consonants were shown in [3]. The study by [4] also showed that some consonants are identified more easily than others, and hence some consonants require a higher SNR in order to be identified as well as others. In addition, the paper also studied the identification of initial vs final consonants. Initial consonants showed a better accuracy of identification as compared to final consonants under all SNR conditions presented. Similar results for initial consonants being more accurately identified were obtained by [5]. One possible argument with a basis in auditory processing could be that a greater neural response for syllable onsets would help perceive them better than syllable offsets [6]. An alternate argument, which has its basis in speech production, is that the initial consonants are produced differently than final consonants (i.e. initial consonants are usually longer and of larger amplitude) [5].

The effect of vowel context on the identification of consonants has been studied previously. It has been shown that consonants are most accurately identified in the presence of the vowel context /a/, next for /i/, and least accurately identified in the context of /u/ [7]. However, these are overall results, and the effect of vowel context can be further studied by separately studying initial vs final consonants or separating consonants based on place of articulation [8]. Previous work has studied the effect of speaking rate on voice-onset time (VOT) and vowel production [9], and on the perception of voice onset time for initial stop consonants [10]. The effect of speaking rate on the identification of consonants in noise, hasn't been studied previously. This paper mainly focuses on this area.

2. Methods

2.1. Perception Experiments

2.1.1. Stimuli and Subjects

The stimuli included only nonsense CVC syllables, since lexical context can affect perceptual results. The CVC syllables were chosen from a set of 120 phonetically balanced nonsense English syllables [3] . A subset of 36 CVCs from the total set, with the corner vowels (/a/, /i/, /u/) were chosen as stimuli for these experiments. Hence the CVCs covered a set of 13 consonants (/p/, /t/, /b/, /d/, /k/, /g/, /s/, /z/, /m/, /n/, /h/, /l/, /r/) and 3 vowels. It is important to note that, in the CVC list that we used, all 13 consonants do not appear in the syllable initial and final location. Table 1 shows the consonants which appear in the initial and/or final position.

The CVC sounds were recorded in a sound proof booth at UCLA using an AKG C-410 head mounted microphone. Two repetitions of each syllable were obtained in the carrier phrase ”/a/-/CVC/ at normal and fast speech rates. This gave a set of 144 CVCs per talker (36 CVCs, 2 repetitions, 2 speaking rates). These were recorded by two native speakers of U.S English. The phrase /a/-/CVC/ was chosen as a compromise between the use of a carrier phrase (which could increase the memory and attention burden for listeners) and using isolated CVCs which lack formant transition cues at the beginning of the syllable. Stimuli were directly digitized at a sampling rate of 16 kHz.

The noisy stimuli were prepared by adding babble noise (from NoiseX database [11]) at an SNR of 0 dB, and 5dB to the clean stimuli. The SNR was calculated using the average SNR of the speech only samples (speech activity detection was performed on the

Table 1: *Consonant List - Initial and Final Syllable Position*

Consonant	**Initial**	**Final**
/p/	✓	✓
/t/	✓	✓
/k/	✓	✓
/b/	✓	
/d/	✓	✓
/g/		✓
/s/	✓	✓
/z/		✓
/h/	✓	
/m/	✓	✓
/n/		✓
/l/	✓	✓
/r/	✓	

clean CVC stimulus). This SNR was used to calculate the noise power to be added to the clean stimulus to prepare the noisy stimulus. Each noisy stimulus was prefixed and postfixed with 75ms of babble noise (at the SNR calculated previously). This was done to enable the listeners to get accustomed to the noise environment. Hence, a total of 864 CVC stimuli (36 CVCs, 2 repetitions, 2 speaking rates - fast, slow, 3 SNRs - quiet, 0dB, 5dB) were presented to each listener.

There were a total of 52 listeners who participated in the experiments. All listeners were normal-hearing adults, less than the age of 25, and were native speakers of U.S. English.

2.1.2. Experimental Setup and Procedure

The experiments were conducted in a sound proof booth at UCLA using the stimuli described in the previous section. The subjects would hear the set of 864 CVC stimuli over two sessions of 1 hour each, comprising of 432 stimuli per session. The stimuli were played back to back, and the subjects were given a 3 second window between successive stimuli, to respond. The subject reported the heard sound by repeating back the stimulus into the microphone. In this setup, the subject didn't need to choose from a list of CVCs. This procedure is similar to that used in [12] and was done for two reasons (a) in order to avoid the errors/bias based on the subjects' knowledge of phonetics, (b) in order to enable the subjects to quickly reproduce the CVC without putting thought into it and causing confusion. A short break of 10 seconds was given after every 20 stimuli in order to avoid fatigue. Two phonetically trained linguists transcribed the responses of the subjects manually. The total number of CVC stimuli transcribed was 864*52 = 44928. The software setup for the listening experiments and the transcriptions were prepared in-house at UCLA. A Matlab based GUI was designed for this purpose.

2.2. Machine Model

2.2.1. Modeling Paradigm

A machine model that is trained to perform phoneme recognition is provided with the same stimuli presented to the human subjects. The machine model we use here is a Gaussian Mixture Model (GMM) based system where a GMM is learned for each consonant (a total of 13). Ideally we would like to split up the set of stimuli into a train and a test set for evaluation. However, we do not have sufficient data to utilize such a framework. There are only 844 stimuli collected, each less than a second long. Hence we train the GMMs using data from the TIMIT corpus which contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. The sentences contain phoneme boundaries. The features used are MFCCs. The GMMs are trained using the Expectation Maximization algorithm (they are not trained discriminatively). The number of mixtures are determined by performing a search on the parameter space. We use 14 mixture GMMs in this work. In the testing phase we find the best phone that matches with each phone segment by performing maximum likelihood on these GMMs. The phone segments are obtained through a manual transcription.

2.2.2. Challenges Faced

- *Mismatched Training and Test Data*: It is important to note that the inherent differences between the training and test corpus make it very difficult to build a good phoneme recognizer. The test corpus consists of *nonsense* consonant-vowel-consonants while the training corpus contains syntactically correct English sentences. In addition, there is a channel mismatch even in clean condition due to different microphone characteristics.
- *Limited Data* : Due to the limited amount of data we are forced to use a paradigm where there is a training/test mismatch. In the future, we hope that a large sized corpus of nonsense Consonant-Vowel-Consonant (CVC) utterances will be collected.

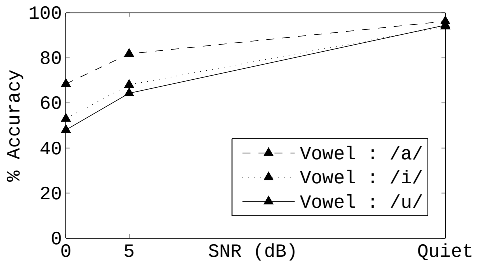
3. Results and Discussion

The results of the perceptual experiments are described in this section. These were analyzed by separating the effect of each of the factors : consonant location (initial or final of a CVC utterance), speaking rate (fast, slow) and vowel context.

3.1. Overall Accuracies separated by Vowel

The overall consonant accuracies separated by vowel are plotted in Figure 1. The highest accuracy is obtained for the vowel context /a/, then /i/ and the lowest for the /u/ context for both speaking rates. This is consistent with the studies performed by [7] and [8].

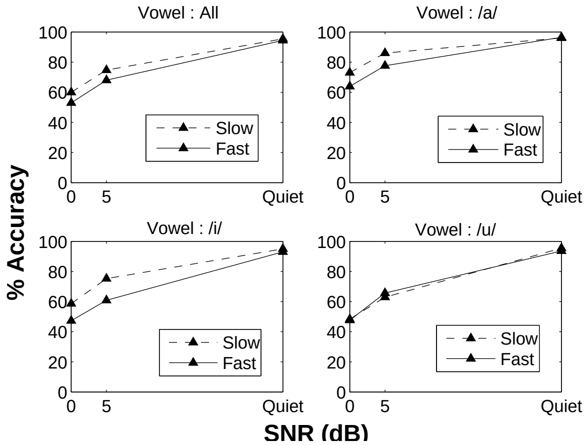
Figure 1: Overall Accuracies (Both Rates)



3.2. Effect of Consonant Location

The consonant position (initial or final in **C**V**C**) is an important factor for the perception of consonants. In Tables 2 and 3, we pick those consonants which occur in both the syllable initial and final positions in our CVC list in the corresponding context (Table 1). The

Figure 2: Effect of Rate



results are reported for the /a/ and /u/ context only since these are least and most affected by noise. We see that the initial consonant is more accurately identified than the final consonants, especially in noise. These results are consistent with previous work of [4] and [5]. The only exception is /l/, where syllable final position was more robust than initial position at both rates. One explanation is that in fast speech, /l/ becomes syllabic and more robust.

Table 2: *Identification accuracies for syllable Initial vs Final Consonants - /a/ context*

Consonant	**Clean**		**5dB**		**0dB**	
	Initial	**Final**	**Initial**	**Final**	**Initial**	**Final**
/t/	0.99	0.93	0.87	0.74	0.66	0.47
/k/	0.97	0.92	0.72	0.45	0.52	0.23
/d/	0.97	0.97	0.95	0.52	0.92	0.44
/l/	0.98	0.96	0.87	0.88	0.61	0.81

Table 3: *Identification accuracies for syllable Initial vs Final Consonants - /u/ context*

Consonant	**Clean**		**5dB**		**0dB**	
	Initial	**Final**	**Initial**	**Final**	**Initial**	**Final**
/p/	0.96	0.66	0.57	0.04	0.42	0.05
/t/	0.98	0.94	0.92	0.68	0.78	0.51
/k/	0.98	0.95	0.54	0.55	0.31	0.34
/s/	0.99	0.96	0.95	0.88	0.93	0.81
/m/	1.00	0.85	0.75	0.23	0.61	0.11

3.3. Effect of Speaking Rate

3.3.1. Separated by Vowel Context

The results shown in Figure 2 are averaged for all consonants and show that humans perceive slow speech more clearly than fast speech in the presence of babble noise for the vowel context of /a/ and /i/. This holds true across both SNR levels that were presented i.e. 0dB, 5dB. However, in the vowel context of /u/, both the slow and fast speech show similar accuracies.

3.3.2. Effect on Voiced Syllable Final Consonants

Table 4 shows the difference in accuracy between slow and fast speech, for voiced and unvoiced consonants. Not all consonants occur in syllable final position (Table 1), and we choose those voiced-unvoiced opposing pairs for which we have complete data. We observe that voiced syllable final consonants are more affected by rate than voiceless consonants in noisy speech. For example, the voiced consonants (/d/ and /z/) are affected more by rate, than their unvoiced counterparts. The duration of the vowel is an important cue for the voicing characteristic of the consonant that follows [13]. A faster speaking rate would shorten the vowel significantly, and it is hence possible that the voicing characteristic of the consonant following it would be hard to identify. We also note from the confusion matrix in Table 9, that the consonant /z/ is usually confused with /s/ and /d/ with /t/. The large difference in intelligibility between the slow and fast speech for voiced consonants could be due to this reason.

Table 4: *Difference in identification accuracies between slow and fast for voiced vs unvoiced syllable final consonants*

		Difference (Slow Accuracy - Fast Accuracy)		
Vowel Context	**Consonant**	**Clean**	**5dB**	**0dB**
/a/	/d/	0.02	0.51	0.33
/a/	/t/	0.05	0.12	-0.11
/i/	/s/	0.02	0	0.01
/i/	/z/	0.16	0.45	0.53

3.3.3. Separated by Consonant Group

In Table 5, the results are separated by consonant groups [Stops, Fricatives, Nasals and Glides]. The identification accuracies of these consonant groups are compared with each other, averaged across all 3 center vowels. The glides (/l/, /r/) are the most robust to noisy conditions, where accuracies as high as 70% are seen in the case of 0dB babble noise. The fricatives show the next best overall performance. We also see the overall trend of slow speech being more intelligible than fast speech, when averaged across all vowel contexts.

Table 5: *Accuracies for each consonant group (all vowels)*

	Clean		**5dB**		**0dB**	
	Slow	**Fast**	**Slow**	**Fast**	**Slow**	**Fast**
Stop	0.96	0.94	0.79	0.59	0.50	0.43
Fricative	0.93	0.92	0.89	0.77	0.76	0.70
Nasal	0.95	0.96	0.55	0.60	0.52	0.43
Glides	0.98	0.97	0.96	0.86	0.70	0.66

3.3.4. Confusion Matrices

Tables 6 and 7 show the confusion matrices for slow and fast speech respectively for the syllable initial consonants, at 0dB SNR. Similarly, Tables 8 and 9 show these confusion matrices for syllable final position. Since the subjects repeat back the identified CVC, and do not choose from a predefined list, it is possible that they report a consonant outside the list of allowed consonants. Those go into the ”none” column of the confusion matrix. Certain rows of the confusion matrix are missing since not all consonants occur in the syllable initial and final position (Table 1). The boldface entries of the confusion matrix which indicate the percentage accuracies for these consonants, are lower for fast speech than slow speech. However, it is evident that the confusion pattern is similar for both speaking rates. For example, /z/ is confused primarily with /s/ in both slow and fast speech. Another example is where /p/ is confused with /k/, /t/ or /h/, for both speaking rates.

In [14], it was reported that /t/ in CV syllables is part of a high scoring set in noise while /b/, /d/, /g/, /k/ are not. We observe a similar trend in our data for both rates and both syllable initial and final position. Similarly, /n/ was reported to be high scoring but not /m/. Here, we again observe a similar trend, however rate affects the perception of /n/ more significantly than /m/. In [14], /s/ and /z/ are both part of the high scoring set, but here we see that while this is true for slow syllable final consonants, /z/ is very low scoring in fast speech due to the reasons outlined in Section 3.3.2.

Table 6: *Confusion Matrix - Slow Speech, 0dB SNR, Syllable Initial Consonants, All Center Vowels*

Slow	**p**	**t**	**k**	**b**	**d**	**g**	**s**	**z**	**h**	**m**	**n**	**l**	**r**	**none**
p	**0.45**	0.11	0.13	0.04	0.01	0.00	0.00	0.00	0.12	0.02	0.00	0.03	0.01	0.06
t	0.14	**0.75**	0.04	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.03
k	0.20	0.05	**0.55**	0.01	0.00	0.00	0.01	0.00	0.15	0.00	0.00	0.00	0.00	0.02
b	0.16	0.01	0.02	**0.60**	0.01	0.01	0.00	0.00	0.09	0.03	0.00	0.01	0.00	0.05
d	0.03	0.10	0.02	0.04	**0.70**	0.02	0.00	0.00	0.04	0.01	0.00	0.00	0.00	0.04
s	0.02	0.02	0.01	0.01	0.00	0.00	**0.90**	0.00	0.01	0.00	0.00	0.00	0.00	0.02
h	0.20	0.02	0.06	0.04	0.00	0.00	0.00	0.00	**0.62**	0.02	0.00	0.00	0.00	0.03
m	0.01	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.05	**0.75**	0.01	0.05	0.03	0.03
l	0.01	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.01	0.23	0.01	**0.57**	0.08	0.05
r	0.03	0.00	0.00	0.14	0.01	0.00	0.00	0.00	0.06	0.10	0.00	0.07	**0.52**	0.05

Table 7: *Confusion Matrix - Fast Speech, 0dB SNR, Syllable Initial Consonants, All Center Vowels*

Fast	**p**	**t**	**k**	**b**	**d**	**g**	**s**	**z**	**h**	**m**	**n**	**l**	**r**	**none**
p	**0.44**	0.09	0.04	0.09	0.01	0.00	0.01	0.00	0.20	0.02	0.00	0.01	0.02	0.08
t	0.13	**0.59**	0.09	0.02	0.02	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.05
k	0.21	0.02	**0.47**	0.02	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.02
b	0.16	0.04	0.04	**0.47**	0.01	0.01	0.00	0.00	0.15	0.03	0.00	0.01	0.01	0.07
d	0.04	0.06	0.01	0.05	**0.74**	0.01	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.02
s	0.00	0.04	0.00	0.00	0.01	0.00	**0.79**	0.01	0.01	0.00	0.00	0.00	0.00	0.13
h	0.14	0.02	0.08	0.02	0.00	0.00	0.00	0.00	**0.70**	0.01	0.00	0.00	0.00	0.03
m	0.03	0.01	0.01	0.05	0.00	0.00	0.00	0.00	0.11	**0.67**	0.01	0.05	0.03	0.03
l	0.03	0.01	0.01	0.04	0.00	0.00	0.00	0.00	0.07	0.16	0.00	**0.57**	0.03	0.07
r	0.01	0.02	0.01	0.03	0.00	0.00	0.00	0.00	0.06	0.06	0.00	0.11	**0.64**	0.06

3.4. Machine Model Results

3.4.1. Overall Consonant Identification Accuracies

The overall accuracies of the machine model are compared to the human results in Table 10. We see that the accuracies of the model are similar to that of the Human at 0dB. The mismatch between the train and test data for the machine model make the clean evaluation files seem ”noisy”. Hence the accuracies of the machine model are far lower than the human results for the clean files.

Table 8: *Confusion Matrix - Slow Speech, 0dB SNR, Syllable Final Consonants, All Center Vowels*

Slow	p	t	k	b	d	g	s	z	h	m	n	l	r	none
p	**0.03**	0.16	0.23	0.01	0.05	0.19	0.02	0.00	0.00	0.00	0.06	0.03	0.00	0.21
t	0.00	**0.45**	0.03	0.00	0.11	0.02	0.01	0.04	0.00	0.01	0.05	0.17	0.00	0.10
k	0.01	0.27	**0.28**	0.00	0.03	0.06	0.02	0.02	0.00	0.02	0.05	0.11	0.00	0.12
d	0.00	0.04	0.00	0.00	**0.50**	0.09	0.00	0.03	0.00	0.01	0.10	0.09	0.00	0.13
g	0.00	0.05	0.01	0.01	0.21	**0.33**	0.01	0.06	0.00	0.02	0.11	0.06	0.00	0.12
s	0.00	0.00	0.00	0.00	0.00	0.00	**0.78**	0.20	0.00	0.00	0.00	0.01	0.00	0.02
z	0.00	0.00	0.00	0.00	0.00	0.00	0.10	**0.87**	0.00	0.00	0.00	0.01	0.00	0.02
m	0.00	0.06	0.00	0.01	0.11	0.07	0.01	0.05	0.00	**0.13**	0.34	0.11	0.00	0.12
n	0.00	0.02	0.00	0.00	0.03	0.02	0.00	0.01	0.00	0.03	**0.68**	0.16	0.01	0.04
l	0.00	0.01	0.00	0.00	0.03	0.00	0.00	0.01	0.00	0.00	0.03	**0.85**	0.00	0.06

Table 9: *Confusion Matrix - Fast Speech, 0dB SNR, Syllable Final Consonants, All Center Vowels*

Fast	p	t	k	b	d	g	s	z	h	m	n	l	r	none
p	**0.06**	0.17	0.55	0.00	0.03	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.13
t	0.00	**0.53**	0.11	0.00	0.01	0.02	0.05	0.04	0.00	0.01	0.06	0.08	0.00	0.10
k	0.01	0.32	**0.34**	0.00	0.01	0.01	0.09	0.03	0.00	0.00	0.03	0.04	0.00	0.11
d	0.00	0.48	0.05	0.00	**0.17**	0.07	0.04	0.03	0.00	0.00	0.05	0.02	0.00	0.09
g	0.01	0.35	0.10	0.00	0.08	**0.22**	0.03	0.03	0.00	0.01	0.03	0.01	0.00	0.13
s	0.00	0.03	0.00	0.00	0.00	0.00	**0.85**	0.07	0.00	0.00	0.00	0.01	0.00	0.03
z	0.00	0.10	0.01	0.00	0.01	0.01	0.49	**0.34**	0.00	0.00	0.00	0.00	0.00	0.03
m	0.01	0.24	0.04	0.00	0.05	0.03	0.04	0.04	0.00	**0.14**	0.25	0.05	0.00	0.10
n	0.00	0.27	0.02	0.00	0.02	0.02	0.03	0.02	0.00	0.01	**0.40**	0.12	0.00	0.08
l	0.00	0.12	0.02	0.00	0.02	0.01	0.02	0.02	0.00	0.01	0.03	**0.69**	0.00	0.07

Table 10: *Machine Model vs Human Accuracies*

	p	t	k	b	d	g	s	z	h	m	n	l	r
Machine - Clean	0.37	0.66	0.88	0.20	0.66	0.07	0.98	0.13	0.68	0.64	0.50	0.46	0.38
Human - Clean	0.89	0.97	0.96	0.98	0.97	0.90	0.90	0.88	0.99	0.95	0.97	0.97	0.99
Human - 5dB	0.49	0.80	0.60	0.68	0.67	0.47	0.86	0.69	0.78	0.60	0.70	0.86	0.90
Human - 0dB	0.35	0.62	0.40	0.54	0.53	0.28	0.82	0.60	0.66	0.45	0.54	0.72	0.58

3.4.2. Effect of Rate

In Figure 3 and Figure 3 we split up the accuracies of by consonant and speaking rate for both the human and machine. Since the files are in clean condition, the accuracy of the human listener is very high(near 98%). The machine performs poorly in clean condition, and hence we try to compare slow/fast for the machine the same way we did for the human subjects in noise. The noisy results for the machine are not presented in this paper. From Figure 3, we see that slow speech has higher accuracy of identification than fast speech for the consonants /p/, /z/, /h/, /m/ , /n/, /l/. The accuracies are similar across both rates for the consonants /g/ and /s/. Fast speech is identified more accurately for the remaining consonants. Hence the overall effect of rate (across all vowels) on the machine model is not very clear. The reason could be that the model that we have used is unable to pick up on certain perceptual cues that the human listeners do.

Figure 3:

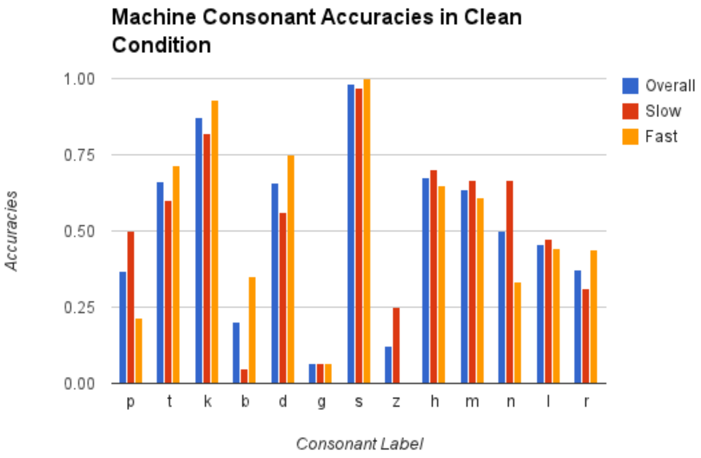
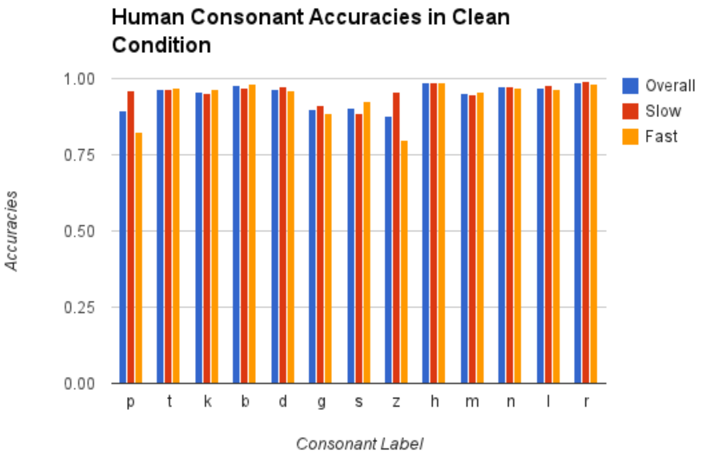


Figure 4:



In Table 11, we split the accuracies by vowel context as well. We notice that overall, the accuracy is the best for the /a/ context, which is the same as the result obtained for that of the human listeners! The effect of rate in /i/ and /u/ context is in complete agreement with that of human listeners (/i/ is better for slow speech, /u/ is invariant to effect of rate). However for the /a/ context, slow speech is better perceived which is different from that of human results. The model has not picked this up.

Table 11: *Machine Model vs Human Accuracies*

Vowel	**Slow**	**Fast**	**Trend Similar To Humans?**
/a/	62.22	68.75	No
/i/	51.58	47.87	Yes
/u/	54.17	54.26	Yes

3.4.3. Confusion Matrices

Tables 12, 13 show the machine confusion matrices for slow and fast speech respectively. An interesting observation is that the machine confusion matrix in clean condition is very similar to the confusion matrix of the human listeners in 0dB babble noise. As mentioned previously, the mismatched training/test condition of the model is analogous to adding noise. From Table 13, we see that /p/ is mainly confused with /t/ and /k/. /t/ is confused primarily with /k/. /s/ and /z/ are confused with each other mainly. For fast speech (Table 13, /z/ is always identified wrongly as /s/. In slow speech the performance is slightly better at 25%. These are in agreement with the confusion matrices of human listeners (Tables 6, etc..)

Table 12: *Machine Confusion Matrix - Slow Speech, All center vowels*

Slow	**p**	**t**	**k**	**b**	**d**	**g**	**s**	**z**	**h**	**m**	**n**	**l**	**r**	**none**
p	**0.50**	0.25	0.06	0.00	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.00	0.00	0.00
t	0.00	**0.60**	0.32	0.00	0.00	0.00	0.04	0.04	0.00	0.00	0.00	0.00	0.00	0.00
k	0.04	0.00	**0.82**	0.00	0.00	0.00	0.07	0.07	0.00	0.00	0.00	0.00	0.00	0.00
b	0.50	0.10	0.00	**0.05**	0.30	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00
d	0.00	0.25	0.06	0.13	**0.56**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
g	0.00	0.00	0.60	0.20	0.07	**0.07**	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
s	0.00	0.00	0.00	0.00	0.00	0.00	**0.97**	0.03	0.00	0.00	0.00	0.00	0.00	0.00
z	0.00	0.00	0.00	0.00	0.00	0.00	0.75	**0.25**	0.00	0.00	0.00	0.00	0.00	0.00
h	0.25	0.00	0.00	0.00	0.00	0.00	0.05	0.00	**0.70**	0.00	0.00	0.00	0.00	0.00
m	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.00	**0.67**	0.12	0.00	0.03	0.00
n	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.08	**0.67**	0.00	0.00	0.00
l	0.00	0.03	0.14	0.00	0.03	0.00	0.00	0.06	0.03	0.08	0.17	**0.47**	0.00	0.00
r	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.19	**0.31**	0.00

Table 13: *Machine Confusion Matrix - Fast Speech, All center vowels*

Fast	**p**	**t**	**k**	**b**	**d**	**g**	**s**	**z**	**h**	**m**	**n**	**l**	**r**	**none**
p	**0.21**	0.50	0.21	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
t	0.04	**0.71**	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
k	0.00	0.07	**0.93**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
b	0.20	0.20	0.00	**0.35**	0.20	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00
d	0.00	0.13	0.13	0.00	**0.75**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
g	0.00	0.00	0.53	0.20	0.13	**0.07**	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
s	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.0 0	0.00	0.00	0.00	0.00
z	0.00	0.00	0.00	0.00	0.00	0.00	1.00	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00
h	0.15	0.00	0.20	0.00	0.00	0.00	0.00	0.00	**0.65**	0.00	0.00	0.00	0.00	0.00
m	0.03	0.00	0.00	0.08	0.00	0.00	0.00	0.14	0.03	**0.61**	0.11	0.00	0.00	0.00
n	0.00	0.25	0.00	0.00	0.17	0.00	0.00	0.17	0.08	0.00	**0.33**	0.00	0.00	0.00
l	0.11	0.08	0.14	0.06	0.03	0.00	0.00	0.06	0.00	0.06	0.03	**0.44**	0.00	0.00
r	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.19	0.00	0.06	0.25	**0.44**	0.00

4. Summary and Conclusion

We have studied the effect of rate on the human perception of consonants in babble noise. This was done by conducting perceptual experiment on human subjects. The data collected from these perceptual experiments are presented and analyzed in this paper. In addition, we also present a modeling study on the data collected. We perform phoneme recognition with the help of a GMM model on clean data. They key results of this paper include : *(a) The intelligibility of syllable final consonants is more affected by noise than initial consonants. (b) Slow speech is better perceived than fast speech for /CaC/ and /CiC/ stimuli. (c) The effect of speaking rate is more pronounced in voiced syllable final consonants than their unvoiced counterparts. (d) The machine in clean condition has a similar effect of rate as the human in /i/ (slow better) and /u/ (rate invariant) context in noise. However, in /a/ context, the model gives a different trend from that of the human (e) The machine in clean condition has similar confusions as that of the human in noisy conditions (0dB).* In future, we will study other consonants and vowel contexts with a much larger set of CVCs. Moreover, we will attempt to incorporate the effect of speaking rate into the model since our model does not replicate the human performance for the /a/ context. We would also work on improving the overall accuracies of the model.

5. References

[1] G. A. Miller and P. E. Nicely, “An analysis of perceptual confusions among some english consonants,” *The Journal of the Acoustical Society of America*, vol. 27, p. 338, 1955.

[2] J. B. Allen, “Articulation and intelligibility,” *Synthesis Lectures on Speech and Audio Processing*, vol. 1, no. 1, pp. 1–124, 2005.

[3] A. Boothroyd and S. Nittrouer, “Mathematical treatment of context effects in phoneme and word recognition,” *The Journal of the Acoustical Society of America*, vol. 84, p. 101, 1988.

[4] J. R. Benkí, “Analysis of english nonsense syllable recognition in noise,” *Phonetica*, vol. 60, no. 2, pp. 129–157, 2003.

[5] M. A. Redford and R. L. Diehl, “The relative perceptual distinctiveness of initial and final consonants in cvc syllables,” *The Journal of the Acoustical Society of America*, vol. 106, p. 1555, 1999.

[6] B. Delgutte and N. Y. Kiang, “Speech coding in the auditory nerve: Iv. sounds with consonant-like dynamic characteristics,” *The Journal of the Acoustical Society of America*, vol. 75, p. 897, 1984.

[7] J. J. Hant and A. Alwan, “A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise,” *Speech Communication*, vol. 40, no. 3, pp. 291–313, 2003.

[8] J. R. Dubno and H. Levitt, “Predicting consonant confusions from acoustic analysis,” *The Journal of the Acoustical Society of America*, vol. 69, no. 1, pp. 249–261, 1981.

[9] R. H. Kessinger and S. E. Blumstein, “Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies,” *Journal of Phonetics*, vol. 26, no. 2, pp. 117–128, 1998.

[10] J. A. Utman, “Effects of local speaking rate context on the perception of voice-onset time in initial stop consonants,” *The Journal of the Acoustical Society of America*, vol. 103, no. 3, pp. 1640–1653, 1998.

[11] A. Varga and H. J. Steeneken, “Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[12] D. L. Woods, E. W. Yund, T. J. Herron, and M. A. U. Cruadhlaoich, “Consonant identification in consonant-vowel-consonant syllables in speech-spectrum noise,” *The Journal of the Acoustical Society of America*, vol. 127, no. 3, pp. 1609–1623, 2010.

[13] L. J. Raphael, “Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in american english,” *The Journal of the Acoustical Society of America*, vol. 51, p. 1296, 1972.

[14] S. A. Phatak and J. B. Allen, “Consonant and vowel confusions in speech-weighted noisea),” *The Journal of the Acoustical Society of America*, vol. 121, no. 4, pp. 2312–2326, 2007.