Truncation Confusion Patterns in Onset Consonants

Andrew Lovitt [a] [b]

IDIAP–RR 07-05

March 2007

soumis à publication

[a] Department of Electrical and Computer Engineering, University of Illinois, Urbana, Il., USA

[b] Institut Dalle Molle d'Intelligence Artificielle Perceptive, Martigny, Valais, Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny − Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

Rapport de recherche de l'IDIAP 07-05

Truncation Confusion Patterns in Onset Consonants

Andrew Lovitt

March 2007

soumis à publication

Résumé. Confusion matrices and truncation experiments have long been a part of psychoacoustic experimentation. However confusion matrices are seldom used to analyze truncation experiments. A truncation experiment was conducted and the confusion patterns were analyzed for 6 consonant-vowels (CVs). The confusion patterns show significant structure as the CV is truncated from the onset of the consonant. These confusions show correlations with both articulatory, acoustic features, and other related CVs. These confusions patterns are shown and explored as they relate to human speech recognition.

1 Introduction

Confusion matrices have long been a part of psychoacoustic phoneme experimentation. Unfortunately truncation experimentation is rarely analyzed based on the confusion patterns. An example of confusion pattern analysis is seen in the 1955 article by Miller and Nicely [1]. The authors analyzed the responses to consonant-vowels (CV) in the presence of variable white background noise was increased. The authors concluded that only certain responses were possible for a given consonant-vowel (CV) at different SNRs. Additionally they found that all consonant's P_c(SNR) did not degrade at the same rate nor did different articulatory features degrade at the same rate versus SNR. This work was later reanalyzed by Soli *et al.* [2] and by Allen [3].

The most famous truncation experiment was conducted by Furui in 1986 [4]. Furui found that in Japanese CVs the $P_c(time)$ changed most dramatically when the cepstrum coefficients changed the most (near transitions). However no mention is made of the confusion patterns in the experiment.

This work is a direct extension of the confusion matrix analysis technique pioneered by Miller and Nicely and then reconstituted by Allen. The experiment is inspired by the work of Furui and analyzed in a way to understand the confusion patterns of the data. The results are analyzed based on the individual utterance confusion patterns and these patterns are analyzed in relation to general trends in the data.

2 Experimental Procedures

The experiment consisted of presenting manipulated CVs to a listener, who then reported what they heard. The CVs were manipulated by truncating the CV in time from the beginning of the consonant. The listeners heard the presentations over headphones[1] in a sound-proof booth. The entire experiment was conducted at two signal-to-noise ratios (SNRs) to ensure the artifacts from truncation were eliminated. The SNRs were 12 (background noise was very quiet) and 0 dB SNR (noticeable background noise). Both were speech-weighted noise[2].

The CVs truncated consisted of the consonants /p/, /t/, /s/, /ʃ/, /z/, and /ʒ/ followed by the vowel of /ɑ/. Each CV was spoken by 10 talkers and was taken from the LDC nonsense speech corpus.[3] Each of the 18 talkers selected from the LDC nonsense speech corpus database spoke either 3 or 4 utterances. All utterances had 16 kHz sampling rate. The utterances taken from the LDC corpus are were proved to have high P_c(0 dB SNR) in speech weighted noise. Thus all the utterances were verified to be well articulated and properly labeled in a preliminary study.

There were 8 listeners from the area surrounding the University of Illinois at Urbana-Champaign. They all spoke English as a first language and had no ear infections or hearing problems. The subjects were trained for an hour by listening to presentations with no masking noise added and no time truncation implemented. This familiarized the subjects with the experiment and the procedure. In the experiment, all presentations were presented randomly to subjects.

Listeners were allowed to choose from any consonant or semi-vowel represented in the LDC corpus. They were also offered a button denoted 'vowel only'. The subject was instructed to only press this button if they repeated the sound more than once and could not hear any consonant. If the listeners heard any consonant they were instructed to make their best guess. The listeners were allowed to proceed at their own pace.

Stimuli Truncation The speech was high pass filtered at 100 Hz to eliminate any DC component. The masking noise was created which had a 12 dB slope below 100 Hz and had a -30 dB slope above

[1] Sennheisser HD 280 Pro

[2] The noise used is the exact same as the noise used in Phatak06 [5]

[3] LDC Articulation Index database number LDC2005S2. For further analysis of this corpus see Fousek04 [6], and Phatak07 [5]

1000 Hz. The sound level and SNR were set analogous to Miller-Nicely [1][4]. The SNR was calculated and the noise was created before any manipulation was performed.

The beginning of the consonant and the beginning of the vowel were found by hand using a combination of spectrograms and raw signal plots. These values were then recorded as the first and last truncation conditions. The stimuli were created by applying a 5 ms ramp[5] immediately before the truncation point. The utterance is truncated to the beginning of the ramp and 500 ms of silence was appended before and after the utterance. Noise was then added to the manipulated utterance and the stimuli was played to the listener. Each truncation condition was exactly 5 ms later than the previous one.

2.1 $P_{h|u}(time)$

Confusion pattern analysis [3], which is a graphical representation of entries in a row of CM plotted as a function of the conditions, is employed to illustrate the results denoted as a $P_{h|u}(time)$ functions. A sample $P_{h|u}(time)$ function is shown in fig. 1. The responses to a utterance u, in fig. 1 $u =$ /sɑ/ for talker f101, are plotted against the truncation condition. The plot's abscissa are normalized to the beginning of the vowel. In this experiment there was no significant difference in the $P_{h|u}(time)$ functions for the 12 and 0 dB SNR conditions of all utterances so all results are collapsed across SNR.

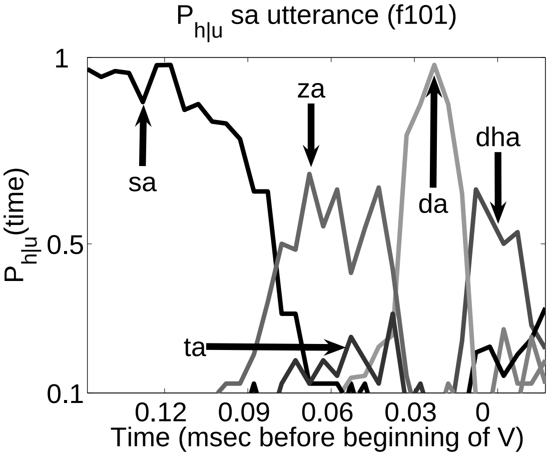

FIG. 1 – This is the $P_{h|u}(time)$ function for /sɑ/ spoken by talker f101. The abscissa shows the amount of time before the onset of vocalic oscillations. The major curves are labeled to aid in understanding the confusions as they relate to the truncation condition relative to the vowel.

After the experiment was completed the onset of the vowel (*vocalic onset* or *vocalic oscillations*) and offset of the consonant energy were manually extracted to aid in the analysis. This was done by using a spectrogram[6] and the plot of the actual waveform. The onset of regular clear vocalic oscillations was taken as the beginning of the vowel. If the consonant was voiced the offset of steady consonant energy was used to help identify the onset of the vowel. These points are used in the analysis to properly align these significant acoustic events with the confusion patterns.

Each curve in the plot corresponds to responses of a different CV. In the example, as the CV is truncated the utterance changes recognition from /sɑ/ to /zɑ/ to /dɑ/ and finally to /ðɑ/ as the truncation reaches the vowel. If the utterance is continually truncated well into the vowel the responses would go to 'only vowel'. This was verified in a preliminary experiment. The following notation will be used in place of the IPA symbols on the graphical results : /θ/ is /th/, /ʃ/ is /sh/, /ð/ is /dh/, /ʒ/ is /zh/, /ʧ/ is /ch/, and /ʤ/ is /j/.

[4] For further information on the software VU meter used, see [7]

[5] The ramp was 1/2 of a hamming window.

[6] 128 samples, overlap of 4, sampling rate 16000 Khz

3 Consonant Confusion Patterns

The 6 CVs tested are grouped into 3 groups based on their confusion patterns. The three groups are /tɑ/ and /pɑ/, /sɑ/ and /zɑ/, and /ʃɑ/ and /ʒɑ/. Each group is analyzed separately because the confusion patterns within each group are related.

3.1 /tɑ/ and /pɑ/

The /tɑ/ and /pɑ/ group have the simplest confusion patterns of all three groups. In this group /tɑ/ utterances are recognized as /pɑ/ when the utterance is truncated beyond a certain critical truncation point. Figure 2 shows both a /tɑ/ and /pɑ/ utterance from the experiment. The /tɑ/ utterance is recognized as /tɑ/ with 100% probability until a range of 2 truncation conditions. These truncation conditions are 15-20 ms after the onset of the burst for the /tɑ/. In all /tɑ/ utterances this critical truncation point is between 10 and 30 ms after the onset of the burst. However all /tɑ/ utterances all have different critical truncation points relative to the beginning of the vowel. This point is between 30 ms and 150 ms before the onset of the vowel. After the critical truncation point the recognition changes to /pɑ/ the recognition is /pɑ/ until the onset of the vowel. All /tɑ/ utterances tested follow this pattern (/bɑ/ and 'vowel only' are slight responses near the vowel).

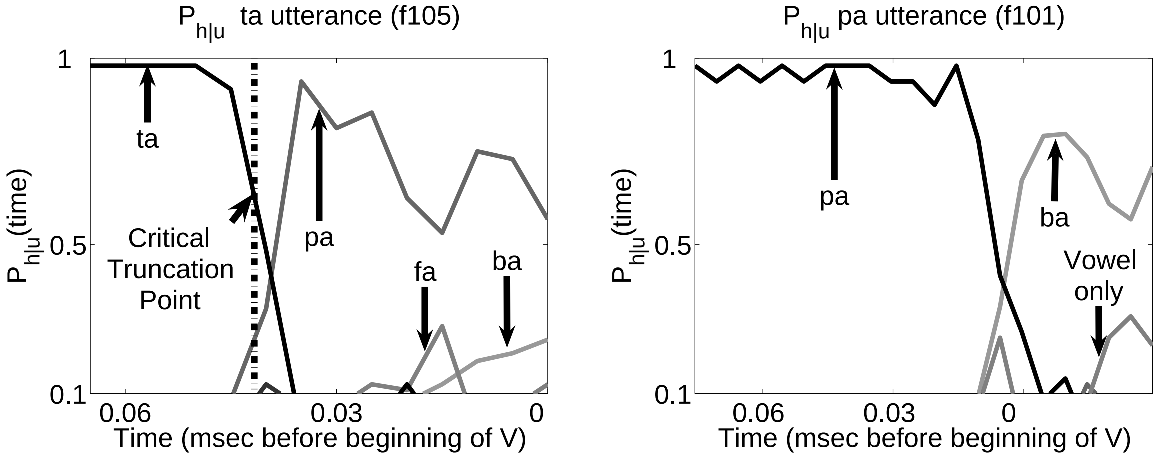

FIG. 2 – The left figure shows a /tɑ/ utterance and the right figure shows a /pɑ/ utterance from the truncation data. The /tɑ/ utterance after the loss of recognition of /tɑ/ and the /pɑ/ utterance have similar confusion patterns up to the onset of the vowel.

The /pɑ/ utterance shows that /pɑ/ is highly recognized until the onset of the vowel. At this point the recognition goes to /bɑ/ and 'vowel only' as the utterance is truncated into the vowel. This confusion pattern is indicative of all /pɑ/ utterances. The /pɑ/ confusion pattern looks like the /tɑ/ recognition pattern after the /tɑ/ critical truncation point. This is similar for all utterances of /pɑ/.

Both of these consonants are unvoiced plosives and the confusion patterns show responses that are almost exclusively plosives and the majority are unvoiced plosives until the onset of the vowel. The /tɑ/ and /pɑ/ relationship is also seen in the confusion patterns seen in a Miller-Nicely repeat experiment. This is reported by Lovitt in 2006 [8].

3.2 /sɑ/ and /zɑ/

The phonemes /sɑ/ and /zɑ/ show a relationship very similar the relationship between to /tɑ/ and /pɑ/ where as one utterance is truncated the confusion patterns of the pair of utterances look similar. Figure 3 shows an example of a /sɑ/ utterances and an example of a /zɑ/ utterances.

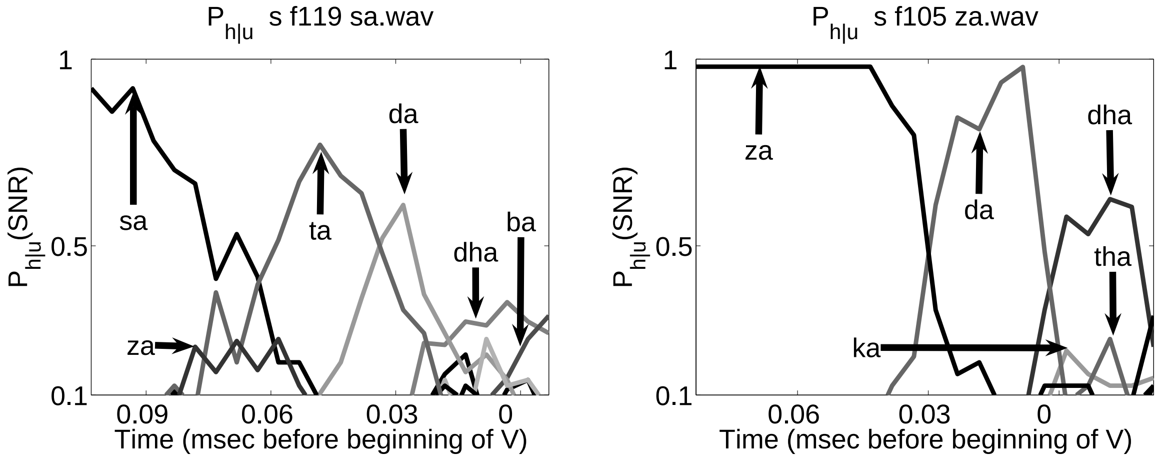

Fig. 3 – The left $P_{h|u}(time)$ is an example that shows a /sɑ/ utterance. This utterance illustrates, with fig. 1, two different confusion patterns seen in the /sɑ/ utterances. The right $P_{h|u}(time)$ shows the results of a /zɑ/ utterances. The /zɑ/ utterances are more homogeneous across talkers than /sɑ/ utterances.

The /sɑ/ utterances (fig. 1 and fig. 3) show the two different confusion patterns for /sɑ/. The first plot (1) shows a /sɑ/ utterance which is recognized as /zɑ/ once the utterance is truncated. The second plot (3) shows a /sɑ/ utterance which is recognized as /tɑ/ instead of /zɑ/ as the utterance is truncated. The recognition of /zɑ/ and the recognition of /tɑ/ happen at approximately the same time relative to the end of the consonant. Each utterance is either reported as confused with /tɑ/ or /zɑ/ as the utterance is truncated (utterances that are confused with both a small minority (1/10)). All /sɑ/ and /zɑ/ utterances are recognized as /dɑ/ after /zɑ/ or /tɑ/. Then they all are confused with /θɑ/, /ðɑ/, /bɑ/, and other consonants as the onset of the vowel is approached. The /zɑ/ utterances all show a progression from /zɑ/ to /dɑ/ to /θɑ/, /bɑ/, /pɑ/, etc. as they are truncated.

There is a strong relationship between /sɑ/ and /zɑ/. Both /sɑ/ and /zɑ/ are confused with the same consonants as the utterance is truncated. Additionally /zɑ/ is only confused with voiced CVs whereas /sɑ/ is confused with both voiced and unvoiced CVs. However the CV /sɑ/ is confused with the same consonants based on articulatory features as /zɑ/ except for the voicing (/tɑ/ and /dɑ/ are the same except for voicing as are /ðɑ/ and /θɑ/). If /tɑ/ and /dɑ/ are collapsed together and /θɑ/ and /ðɑ/ are collapsed together the confusion patterns are extremely similar. Also all major responses for both /sɑ/ and /zɑ/ (/sɑ/, /zɑ/, /tɑ/, /dɑ/, /θɑ/, and /ðɑ/) have the same place articulatory feature according to Miller-Nicely 1955. These confusion between /sɑ/ and /zɑ/ (and the robustness of place features) is another example of a confusion pattern seen in white masking noise as the noise is raised [8].

Figure 4 shows the probability of certain responses given the time relative to the end of the consonant for /sɑ/ and /zɑ/. The confusions correlate with offset of the consonant. The left figure in fig. 4 shows the average responses for /sɑ/ and the right figure in fig. 4 is for /zɑ/. In both plots the black dashed line shows the probability of either /sɑ/, /zɑ/, /tɑ/, /dɑ/, /ðɑ/, /θɑ/, or /bɑ/. Each CV is represented by an unique line type. This set of consonants describes well the possible responses up to the offset of the consonant and describes the structure of the responses after that point well, though with less probability. The offset of the consonant is shown as a red dashed vertical line. The reason /sɑ/ doesn't reach 100% in slightly truncated conditions is that one listener confused /sɑ/ with /kɑ/ so that /sɑ/ and /kɑ/ were reported 50% of the time each in that region for all /sɑ/s.

The placement of each peak in recognition for each consonant may be 5 or 10 ms different on a per utterance basis but the average represents the data well. The shape of /bɑ/ and /ðɑ/ and /θɑ/ show very similar patterns for both /sɑ/ and /zɑ/ in placement and probability of response. The responses which were /dɑ/ when /zɑ/ is truncated approximate the shape and timing of /tɑ/ and /dɑ/ when

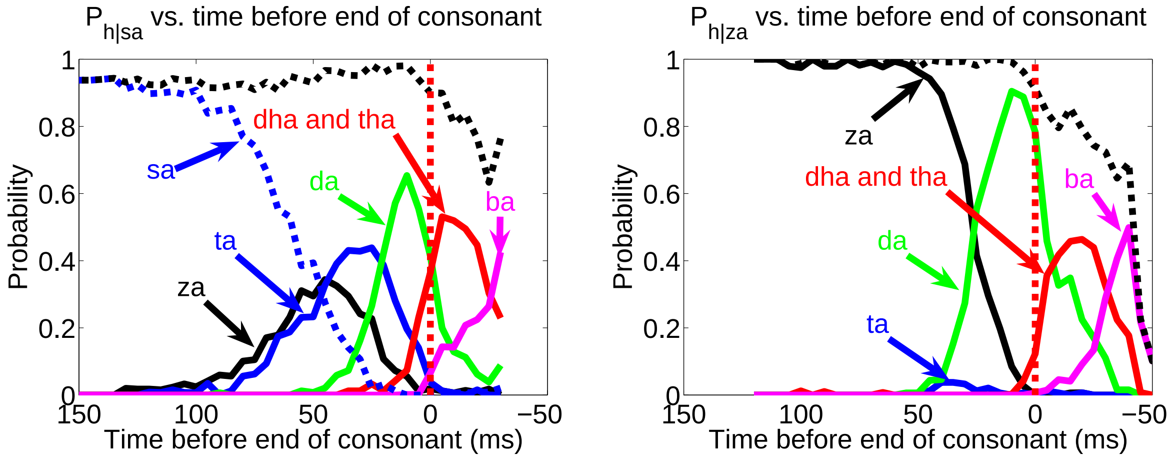

Fig. 4 – These figures show the distribution of responses for /sɑ/ (left) and /zɑ/ (right) based on the time truncation relative to the end of the burst in the consonant. Both show very similar patterns with similar probabilities. Additionally the competition between /zɑ/ and /tɑ/ is seen as both curves in the /sɑ/ case are extremely overlapping.

added together from /sɑ/ truncations.

These results show a strong relationship between the confusion patterns for /sɑ/ and /zɑ/. Both /sɑ/ and /zɑ/ have the same articulatory features except for voicing. However once either CV is truncated to 25 ms before the offset of the burst they have almost identical patterns. Additionally the confusion of /sɑ/ with /zɑ/ shows a strong connection between /sɑ/ and /zɑ/ in recognition which is supported by other experiments [8].

3.3 /ʃɑ/ and /ʒɑ/

In this group there are confusions with affricated consonants. In fact both /ʃɑ/ and /ʒɑ/ have a high number of affricated responses as the utterance is truncated. Figure 5 shows an utterance of both /ʃɑ/ and /ʒɑ/.

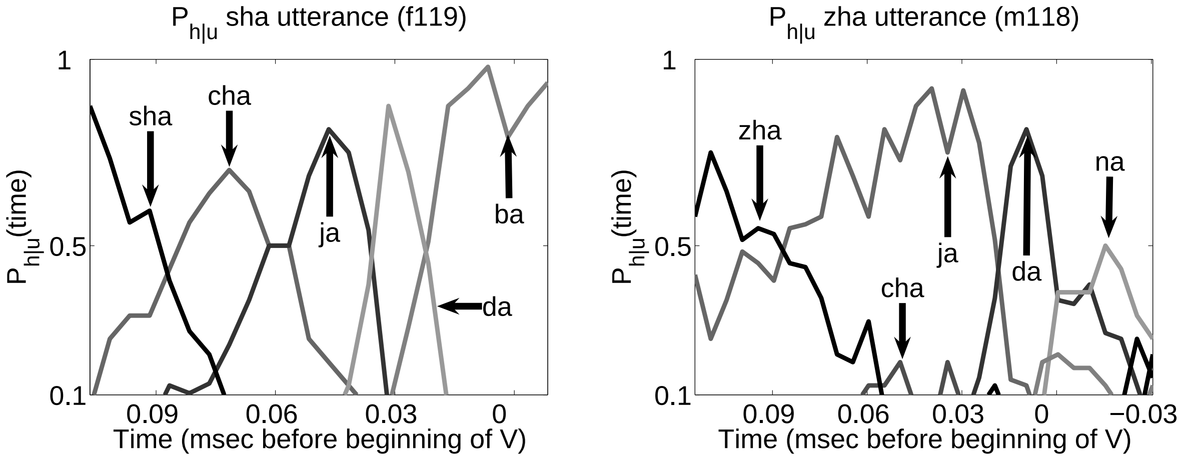

Fig. 5 – These figures show confusion patterns for /ʃɑ/ (left) and for /ʒɑ/ (right). These utterances show confusions with affricates and plosives, the later is similar to the confusions seen in /sɑ/ and /zɑ/.

The /ʃɑ/ utterance shows a /ʃɑ/ which is confused with /ʧɑ/ and then a /ʤɑ/ utterance as the utterance is truncated. This utterance then proceeds to /dɑ/ and then to /bɑ/ as the truncation approaches the vowel. The large confusions with /ʧɑ/ is seen in all /ʃɑ/ truncated utterances. These /ʧɑ/ responses at are truncation conditions where the truncation artificially introduces a strong onset synchrony across all frequencies in the consonant and the frication is long in duration (> 20 ms). This is mirrored in the confusions of /ʤɑ/ with /ʒɑ/. This strong onset synchrony is seen in every /ʃɑ/ and /ʒɑ/ utterance as they are truncated. the difference between /ʧɑ/ and /ʤɑ/ is voicing which is reminiscent of /tɑ/ and /dɑ/ confusions in /sɑ/ and /zɑ/.

The /ʒɑ/ utterances show some differences. The /ʒɑ/ confusion patterns show exclusive confusion with voiced consonants, just like /zɑ/ confusions. The recognition of /ʒɑ/ is not more than 50% in most utterances when there is no truncation. This is due to the listeners not being familiar with the /ʒɑ/ phone. The CV /ʒɑ/ is very rare, /ʒ/ is also rare, in English when compared with /ʤɑ/. Because of this, /ʒɑ/ was significantly confused with /ʤɑ/.

The similarities in the types of confusions also extend to the average confusion patterns. Figure 6 shows the average confusion patterns for /ʃɑ/ and /ʒɑ/ are similar to the patterns found in fig. 4 for /sɑ/ and /zɑ/. In both cases the recognition of /dɑ/ and /bɑ/ is very high near the end of the burst and the beginning of the vowel respectively. The combined /ʧɑ/ and /ʤɑ/ responses for /ʃɑ/ are similar to the /ʤɑ/ confusions in /ʒɑ/.

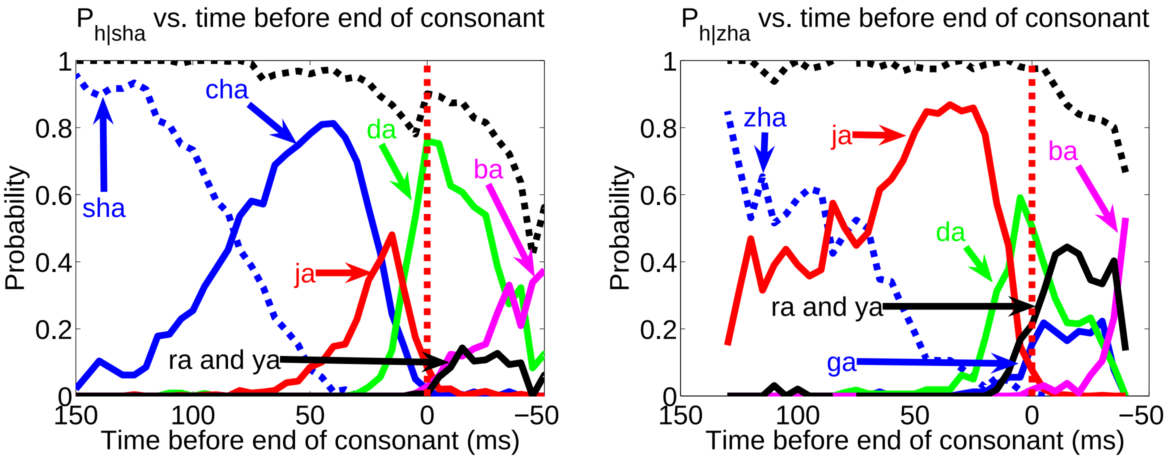

FIG. 6 – These figures show the average confusion patterns for /ʃɑ/ (left) and /ʒɑ/ (right). As with the figures in fig. 4 the structure shows the lowering of performance from the actual spoken CV and confusions which are systematic based on the truncation location relative to the end of the consonant.

4 General Trends

There is a large amount of structure seen in the comparison of the responses between the CVs. Both /ʃɑ/ and /ʒɑ/ have a large frequency range for the burst of the consonant. This range is similar to the range for the burst in affricates. The differences in the signals as previously stated relates to the strong onset synchrony in the burst created by the truncation. The length of the burst seems to be correlated well with the confusions. This is seen also in the comparison between /sɑ/ and /zɑ/. There is a correlation between the minimum length of the burst and /sɑ/ or /zɑ/ responses. The /sɑ/ responses begin to be confused with /zɑ/ before the point where the /zɑ/ responses are confused with other CVs. Thus there is a small region where the length of the burst is too short for /sɑ/ but long enough to be /zɑ/. This length is between 25 and 75 ms. before the end of the consonant.

There is a similar analysis that can be seen for the responses to /tɑ/ and /dɑ/. The /tɑ/ responses for /sɑ/ are only when the burst is longer than the /dɑ/ response conditions. This is due to the relative position between the burst and the onset of the vowel. In /sɑ/, the only CV to show this /tɑ/

effect, the time between the offset of the burst and the onset of the vowel is between 10 to 40 ms. Whereas in /zɑ/ the time is between 0 and 10 ms. This relative timing has an effect on whether /tɑ/ is a possible response or all responses are /dɑ/ when the frication is short.

The results from the analysis of all the truncation confusion patterns tested lead to the concept of a hierarchy of consonants in time. All CVs in this experiment are recognized as /bɑ/, /pɑ/, and /dɑ/ near the vowel or near the end of frication. These responses are related to the original CVs by either the lengthening of a feature (for instance frication time) or the addition of an event (for instance a burst as in /tɑ/). Figure 7 shows a cartoon of the flow of recognition as an utterance is truncated. This cartoon is meant to illustrate the trends and is not plotted data..

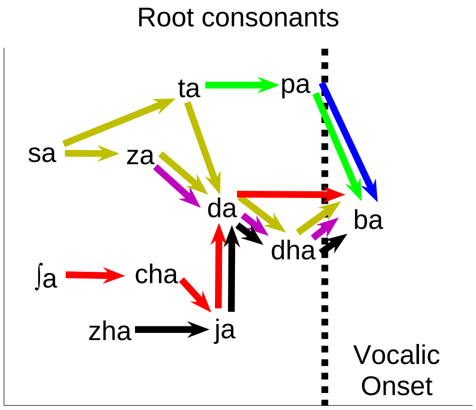

Fig. 7 – This cartoon shows the progression of confusable phones as CVs are truncated from the beginning of the consonant. All the data in the experiment is represented in this cartoon.

An example is the /dɑ/ recognition in all fricatives. The /dɑ/ responses are the largest response right before the end of the burst in the consonant. At this condition the acoustic signal will have a burst with a very short duration immediately proceeding the onset of the vowel. This situation causes the recognition of /dɑ/. However as this burst is lengthened in time (truncation point is moved towards the beginning of the consonant) the recognition changes to an affricate or fricative. This implies that the length of frication causes the different recognition responses.

5 Conclusions

The experiment was conducted to analysis the confusions found in truncation experiments. The experiment provided results that showed that the $P_c(time)$ of the CVs in fact dropped significantly before the transition. This is in opposition of the results from Furui. These confusions shows systematic structure and structure in the articulatory features that compose the responses. The relationship between /ʃɑ/ and /ʧɑ/, and /ʒɑ/ and /ʤɑ/ also illuminates some systematic differences in the CVs and ability of English subjects to understand them (/ʒɑ/ and /ʤɑ/ confusions). This all leads to a more complete picture of the relationship between CVs and the their recognition.

6 Acknowledgments

I would like to thank the HSR research group at UIUC for aiding in the collection of the data for this analysis. I would also like to thank Jont Allen for his advice and funding for data collection. Analysis of the perception results was done under the DIRAC integrated project funded by the European Community.

Références

[1] George A. Miller and Patricia E. Nicely, "An analysis of perceptual confusion amoung English consonants," *Journal of the Acoustical Society of America*, vol. 27, pp. 338–352, 1955.

[2] Sigfrid D. Soli and Phipps Arabie, "Auditory versus phonetic accounts of observed confusions between consonant phonemes," *Journal of the Acoustical Society of America*, vol. 66, pp. 46–59, 1979.

[3] Jont B. Allen, "Consonant recognition and the articulation index," *Journal of the Acoustical Society of America*, vol. 117, pp. 2212–2223, April 2005.

[4] Sadaoki Furui, "On the role of spectral transition for speech perception," *Journal of the Acoustical Society of America*, vol. 80, pp. 1016–1025, 1986.

[5] S. Phatak and Jont B. Allen, "Consonant and vowel confusions in speech-weighted noise," *Journal of the Acoustical Society of America*, Apr. 2007, In press.

[6] Petr Fousek, Petr Svojanovsky, Frantisek Grezl, and Hynek Hermansky, "New nonsense syllables database – analyses and preliminary asr experiments," in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, October 2004, IDIAP-RR 2004-29.

[7] B. Lobdell and Jont B. Allen, "Modeling and using the vu-meter (volume unit meter) with comparisons to root-mean-square speech levels," *Journal of the Acoustical Society of America*, vol. 121, no. 1, pp. 279–285, 2007.

[8] Andrew Lovitt and Jont Allen, "50 years late : Repeating miller-nicely 1955," in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, September 2006.