SPECTROGRAM READERS' IDENTIFICATION OF STOP CONSONANTS

Lori F. Lamel

Department of Electrical Engineering and Computer Science, and
Research Laboratory of Electronics
Massachusetts Institute of Technology, USA

ABSTRACT -- This paper reports on the performance of five spectrogram readers at identifying spectrograms of stop consonants extracted from continuous speech. The stops were spoken by 299 talkers and were presented in the immediate phonemic context. The task was designed to minimize the use of lexical and other higher sources of knowledge. The averaged identification rate across contexts ranged from 73-82\% for the top choice, and 77-93% for the top two choices. The readers' performances were comparable to those of other spectrogram reading experiments reported in the literature, however the other studies have typically evaluated a single subject on speech spoken by a small number of talkers.

INTRODUCTION

Since the invention of the sound spectrograph (Koenig, Dunn, and Lacey, 1946), spectrograms have been widely used by researchers in the speech community. Spectrogram reading involves interpreting the acoustic patterns in the image to determine the spoken utterance. One must selectively attend to many different acoustic cues, interpret their significance in light of other evidence, and make inferences based on information from multiple sources. These include knowledge of the acoustic correlates of speech sounds and their contextual variation, and phonotactic constraints. While early attempts at spectrogram reading met with limited success, the richness of phonetic information in the speech signal was illustrated in a series of experiments (Cole et al., 1980) which demonstrated that high performance phonetic labeling of a spectrogram could be obtained. However, the spectrogram reading experiments reported in the literature have typically evaluated a single spectrogram reader on speech spoken by a small number of talkers. A summary of previous spectrogram reading experiments is given in Lamel (1988). This paper describes spectrogram reading experiments conducted to evaluate the ability of spectrogram readers and to obtain a better understanding of the process. Several experienced spectrogram readers were evaluated on speech from many talkers and in a variety of local phonemic contexts.

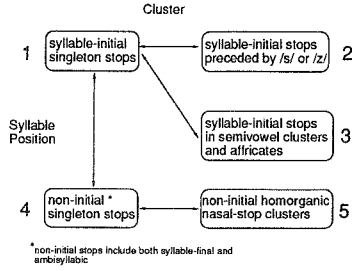


Figure 1: Experimental Design

DESCRIPTION OF THE EXPERIMENTS

The experiments assessed the subjects' ability to identify stop consonants presented in only their immediately surrounding phonetic context. The tokens, extracted from continuous speech, consisted of a stop or a two-consonant sequence containing a stop, and a single vowel on each side. As a class, the stop have been extensively studied; their articulation is complicated, consisting of dynamic characteristics which vary depending on context (e.g., Fant, 1960). Stops are also among the most frequently occurring sounds in English (Denes, 1963), appearing both alone and in a variety of consonant clusters.

Subjects identified stop consonants in five different contexts as shown in Figure 1. The first task assesses the subjects' ability to identify singleton stop consonants in syllable-initial position. After establishing this baseline performance, the effects of intervening consonants and syllable position on the subjects'

decision can be determined. Acoustic studies have shown that the acoustic characteristics of stops in syllable-initial consonant clusters change from the canonical characteristics of singleton stops (Lehiste, 1962; Zue, 1976). The remaining tasks evaluate the subjects' ability to identify stop consonants in clusters with other consonants and in non-syllable-initial position.

The speech tokens were selected from two time-aligned, phonetically-transcribed speech databases (Ice Cream and TIMIT) developed at MIT. A total of 615 tokens, spoken by 299 talkers and extracted from over 500 sentences, were identified. Each token consisted of the stop or consonant sequence and both the preceding and following vowels in their entirety. The tokens were almost equally divided between male and female talkers. The identity of each token was specified by its phonetic transcription. The selected tokens were extracted from digitized continuous speech and randomized. A spectrogram was made of each token using the *Spire* facility (Cyphers, 1985). Example tokens are shown in Figures 5 and 6.

Table 1: Number of readers and tokens for each task

Task number	Number of subjects	Number of tokens	Percent male/female	Number of vowel contexts
1	5	263	52/48	101
2	2	102	46/54	60
3	1	51	41/59	32
4	3	153	57/43	87
5	1	46	59/41	35

Five spectrogram readers participated in the experiments. The experience of the subjects varied; one subject has been reading spectrograms for about 15 years investing over 3000 hours, the other four subjects have been reading spectrograms for 4 to 8 years, estimating their experience in the range of 300 to 700 hours. All of the readers have taught the MIT version of a one-week intensive spectrogram reading course at least three times. Table1 shows the number of subjects and total number of tokens of spectrograms read for each task. Readers identified the consonant as one of the stops /b,d,g,p,t,k/; for task 3 the affricates /j,č/ were also included. Subjects were encouraged to give alternate choices when they were uncertain of their decision.

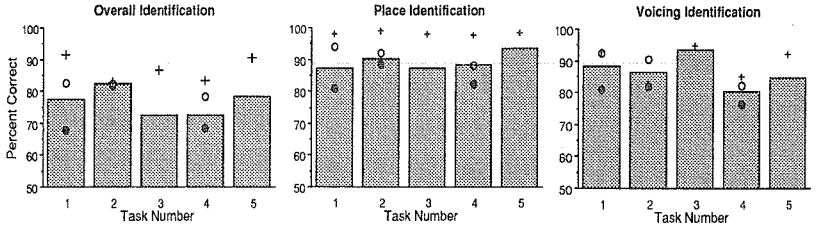


Figure 2: Readers identification rates for each task: overall, place of articulation, and voicing characteristic. The bar represents the average score for all readers. The open circle (°) is the best reader's score, the filled circle (•) shows the worst reader's score, and the + denotes the average listeners' identification.

RESULTS AND DISCUSSION

Figure 2 gives the overall identification rates for each of the five tasks. The bar graph represents the averaged scores for all readers. When multiple readers participated in the experiment, the best and worst accuracies are also plotted. Averaged listener scores are provided for comparison. With the exception of task 2, spectrogram readers identified stops 10-15% less accurately than did the listeners. The average identification rate for singleton, syllable-initial stops was 77.6%, while individual readers' rates ranged from 68-83% correct. The averaged identification rate on task 2, syllable-initial stops preceded by /s/ or /z/, was 82.3%, with a 1% difference between the readers. Unvoiced stops preceded by /z/ had the lowest error rate of 5%. Unvoiced stops preceded by /s/ had about the same error rate (7-8%), whether or not the stop was in a cluster with the /s/. Voiced stops preceded by /s/ had the highest error rate of 40%. The accuracy in task 3, consisting of stops in syllable-initial semivowel clusters and affricates, was 72.5%. Alveolar stops had the largest error rate (over 50%), being confused primarily with the affricates: /dr/ was labelled as /j/ more frequently than it was correctly identified. The affricates were more likely to be confused with each other, than to be called alveolar stops.

Readers identified non-syllable-initial singleton stops about 5% less accurately than syllable-initial singleton stops. The average correct identification in task 4 was 72.5%, with an inter-subject range from 69% to 78%. The subject in task 5 had an accuracy of 78.3% on non-syllable-initial stops in homorganic nasal-stop clusters. This represented an 8% improvement for the reader over task 4, suggesting that the presence of the nasal aided the identification of the stop.

Figure 2 also shows the identification of the place of articulation and the voicing characteristic for all five tasks. Spectrogram readers identified place of articulation 5-10% less accurately than did listeners, with a variation across tasks ranging from 87.1% for tasks 1 and 3 to 93.5% for task 5. Readers' identification of voicing ranged from a low of 80.4% for task 4 to 93.5% for task 5.

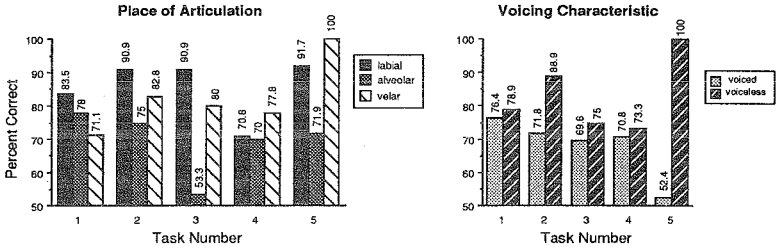


Figure 3: Readers' identification accuracy of stops as a function of the place of articulation and voicing characteristic of the stop.

Figure 3 shows the stop identification accuracy of spectrogram readers as a function of the place of articulation and of the voicing characteristic of the stop. In syllable-initial position, labial stops had the highest identification rate. Labial stops may be easiest to identify, as they are typically weak and have a release that is spread across all frequencies. For the non-initial stops, velars were identified most accurately. The compactness of velars may be a strong cue to their identification. As shown in Figure 3, voiceless stops were always identified more accurately than voiced stops .

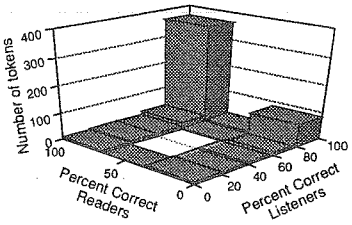


Figure 4: Readers' accuracy as function of listeners' accuracy.

Performance relative to listeners

The spectrogram reader results were correlated with the listeners' performance (Lamel, 1988). As expected, readers labeled tokens that were heard without error more accurately than they labeled tokens that had been misidentified by listeners. Figure 4 shows the accuracy for readers as a function of the accuracy for listeners in the form of a three-dimensional histogram. There are only three values on the reader axis: 0%, 50% and 100%. This is because each token was read by at most two readers. Most tokens were read and heard correctly. The tokens that were read incorrectly often were heard incorrectly too. The errors made by the readers agreed with at least some listener in about 70% of the cases. The tendency for readers to make errors similar to those made by listeners suggests that spectrogram readers may be extracting perceptually relevant information when performing the labeling task. The performance of the spectrogram readers was consistently worse than that of the listeners. In particular, readers were worse at identifying place of articulation than were listeners. Listeners always had 98% or better place identification. There are several possibilities including the spectrographic representation, our inability to

identify and locate acoustic attributes in the spectrogram, and our inability to deduce the phonemes from the attributes. The relative importance of these factors is difficult to assess.

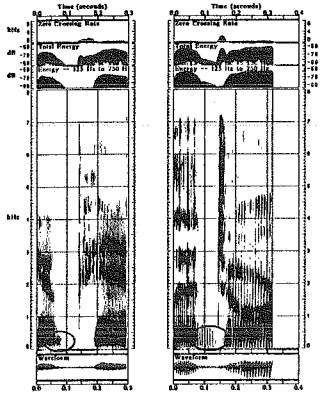
Alternate choices

Table 2 shows the number of cases in which readers supplied an alternate choice. The alternate choices provide an indication about which features were in question. Since the readers choices usually differed in either place or voicing, a partial feature specification could be provided by considering both alternatives. With the exception of task 3, where the indecision was between affricates and alveolar stops, the second choices were almost evenly divided between place and voicing.

Table 2: Readers' responses when alternative choices were supplied.

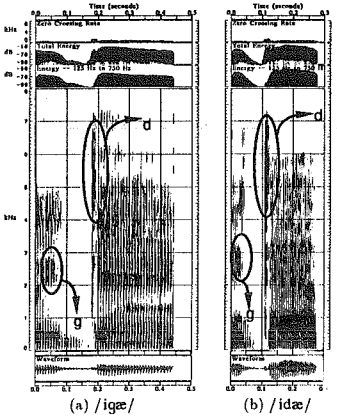
Task number	Number of multiple choices	1st choice correct(%)	% correct in top 2	comment
1	99	63	86	50% unsure of voicing
2	58	78	91	91% unsure of voicing
3	10	70	80	40% unsure of voicing 40% alveolar--affricate
4	102	80	93	64% unsure of voicing
5	20	65	100	75% unsure of voicing

Some readers provided alternate choices frequently, while others hardly ever gave them. In over 63% of these cases the top choice was correct. When only one choice was given, the readers' averaged accuracy was 85%. When multiple choices were given, the readers' averaged top choice accuracy was 67.5%. This difference indicates that the readers often knew when they were uncertain. The improvement in accuracy obtained by including the second choices also shows that readers often knew which feature was uncertain.



(a) /upɪ/ (b) /ədo/

Figure 5: Spectrograms with conflicting voicing information.



(a) /igæ/ (b) /idæ/

Figure 6: Spectrograms with conflicting place information.

Spectrogram readers' use of acoustic attributes

While what spectrogram readers do when they interpret a spectrogram can not be definitively concluded, some of the important acoustic attributes can be identified and inferences made as to how they are used. The inferences were based on the markings and comments made on the spectrogram by the readers, by discussion of the labels with the readers, and by introspection.

Spectrogram readers tend to decide voicing and place of articulation independently. The acoustic attributes used for voicing are primarily the VOT, the presence or absence of aspiration, and the presence or absence of prevoicing during closure. For syllable-initial stops (not in /s/-clusters), VOT seems to be the most important cue. However, in noticing whether or not the VOT is short or long, readers are probably

also determining whether or not the stop is aspirated, and using that information simultaneously. When the VOT is medium, readers check for aspiration and for prevoicing. If the stop is clearly aspirated, readers are generally willing to ignore prevoicing. When readers are uncertain about the aspiration, they weigh prevoicing more heavily--- if there is prevoicing, then the stop is more likely to be voiced than voiceless. Readers write comments like ``VOT medium, is this aspiration? --- voicing hard to tell'' on their spectrograms. They also sometimes circle prevoicing to indicate they used that cue in forming their decision. Two examples with conflicting voicing information are shown in Figure 5. The reader circled prevoicing during the closure interval of both tokens. In (a) the stop is clearly aspirated, and the spectrogram reader weighed that information more importantly, correctly identifying the stop as a /p/. Since this reader was particularly conservative, he also proposed a /b/ as a second choice. The stop in (b) has strong prevoicing throughout the closure. The stop has a medium VOT and it is unclear whether or not it is slightly aspirated. The strength of the prevoicing allowed the reader to correctly determine the voicing of the stop.

From the markings made on the spectrograms by readers, it appears that they primarily use the frequency location and distribution of the burst, the burst strength, and the formant transitions to determine the place of articulation of the stop. These three sources of information may either be confirmatory or contradictory. When they are confirmatory, such as for a labial stop that has a weak, diffuse release falling formant transitions, readers are fairly confident in their label. When the information is contradictory, readers are uncertain and tend to weigh one or two factors more heavily, disregarding the contradictory information. Three examples with conflicting information are shown in Figure 6. The stop in (a) is a /g/, the second candidate given by the reader. The readers arrows indicate that he liked the release best as alveolar and the formant motion on the left as velar. In this case, the reader favored the burst location over the formants and misidentified the stop. In (b) the same reader faced the same contradictory information. Here the reader once again favored the burst location over the formants, and this time was correct.

SUMMARY

Experiments were performed to assess human spectrogram readers' ability to label stops in limited phonetic environments. They also represent an effort to better understand some of the factors involved in spectrogram reading. The evidence, obtained from both spectrogram reading experiments and from teaching spectrogram reading, suggests that the process can be modeled with a set of rules. Knowledge obtained from experiments like the ones presented here can aid in out attempts to formalize the spectrogram reading process. Formalizing spectrogram reading entails refining the language (terminology) that is used to describe acoustic events on the spectrogram, and selecting a set of relevant acoustic events that can be used to distinguish phones. Rules are then used to combine these acoustic attributes and to make phonetic judgments The rules need to account for contextual variation (coarticulation), and partial and/or conflicting evidence, and to be able to propose multiple hypotheses.The markings and comments provided by the readers were helpful in understanding which acoustic attributes are used, and how much weight they are given.

These spectrogram reading experiments indicated that:

- Spectrogram readers were able to label stop consonants across a large number of speakers and many phonemic environments with only a limited phonetic context. The accuracy is consistent with other reported studies (Bush, Kopec, and Zue, 1983; Cole and Zue, 1980).

- On the average, readers identified stops 10-15% less accurately than did listeners. The difference in accuracy may be due to our incomplete knowledge of how to detect acoustic attributes in the spectrogram and how to use the attributes to form phonetic hypotheses, and in part due to inadequacies in the spectrographic representation.

- Syllable position and additional consonants affected the readers' ability. Singleton stops were better identified in syllable-initial position than in non-initial position Initial stops preceded by /s/ or /z/ had a slightly higher voicing error rate than did singleton stops. In non-initial position, stops in homorganic nasal clusters were identified better than singleton stops. Readers confused the clusters /dr,tr/ with the affricates /j,č/. These trends are the same as were observed for the listeners.

[This research was supported by DARPA under contract N00014-82-K-0727,monitored through the Office of Naval Research.]

REFERENCES

Bush, M.A., Kopec, G.E. & Zue, V.W. (1983) "Selecting Acoustic Features for Stop Consonant Identification," *Proc. IEEE ICASSP-83*, 742-725.

Cole, R.A., Rudnicky, A.I., Zue, V.W. & Reddy, D.R. (1980) "Speech as Patterns on Paper," Ch. 1 in *Perception and Production of Fluent Speech*, Cole, R.A., ed., (Lawrence Erlbaum: NJ).

Cole, R.A. & Zue, V.W. (1980) "Speech as Eyes See It," Ch. 2 in *Attention and Performance VIII*, Nickerson, R.S., ed., (Lawrence Erlbaum: NJ).

Cyphers, D.S. (1985) *Spire: A Research Tool*, S.M. Thesis, Massachusetts Institute of Technology.

Denes, P.B. (1963) "On the Statistics of Spoken English," *JASA* 35, no. 6, 892-904.

Fant, G. (1960) *Acoustic Theory of Speech Production*, (Mouton: The Hague).

Koenig, W., Dunn, H.K. & Lacey, L.Y. (1946) "The Sound Spectrograph," *JASA* 18, no. 1, 19-49.

Lamel, L.F. (1988) *Formalizing Knowledge used in Spectrogram Reading: Acoustic and perceptual evidence from stops,* Ph.D. Thesis, Massachusetts Institute of Technology.

Lehiste, I. (1962) "Acoustical Characteristics of Selected English Consonants," Report No. 9, U. Michigan, Communication Sciences Laboratory, Ann Arbor, Michigan.

Zue, V.W. (1976) *Acoustic Characteristics of Stop Consonants: A Controlled Study,* Ph.D. Thesis, Massachusetts Institute of Technology.