THE CONTRIBUTION OF CONSONANTS VERSUS VOWELS TO WORD RECOGNITION IN FLUENT SPEECH

Ronald A. Cole *Yonghong Yan* *Brian Mak* *Mark Fanty* *Troy Bailey*

Center for Spoken Language Understanding
Oregon Graduate Institute of Science and Technology
20000 N.W.Walker Road, Portland, OR 97006
{cole, yan, mak, fanty, bailey}@cse.ogi.edu

ABSTRACT

Three perceptual experiments were conducted to test the relative importance of vowels vs. consonants to recognition of fluent speech. Sentences were selected from the TIMIT corpus to obtain approximately equal numbers of vowels and consonants within each sentence and equal durations across the set of sentences. In experiments 1 and 2, subjects listened to (a) unaltered TIMIT sentences; (b) sentences in which all of the vowels were replaced by noise; or (c) sentences in which all of the consonants were replaced by noise. The subjects listened to each sentence five times, and attempted to transcribe what they heard. The results of these experiments show that recognition of words depends more upon vowels than consonants—about twice as many words are recognized when vowels are retained in the speech. The effect was observed when occurrences of [l], [r], [w], [y] [m], [n], were included in the sentences (experiment 1) or replaced by noise (experiment 2). Experiment 3 tested the hypothesis that vowel boundaries contain more information about the neighboring consonants than vice versa.

1. INTRODUCTION

Do vowels or consonants convey more information about words in fluent speech? We address this question by removing information about consonant segments or vowel segments from spoken sentences, and examining the effect on word recognition performance.

The experiments reported here investigate word recognition using read sentences from the TIMIT database[1]. These sentences are a good choice for research on the relationship between phonetic information and word recognition because the speech corpus is in the public domain, the speech is produced by many different speakers, and each utterance is annotated with time-aligned phonetic transcriptions and orthographic word level transcriptions.[1]

For the purposes of this research, we grouped TIMIT labels into three sets of sounds, called **consonants**, **vowels** and **weakson** (weak sonorants, for lack of a better name), as shown in Table 1. The consonant class consists of 20 obstruent consonants. The vowel class consists of vowels and dipthongs. The weakson class consists of liquids ([l], [r], [el]), glides ([w], [y]), and nasals ([m], [n], [nx], [ng], [em], [en], [eng]). Although nasals are classified as +consonantal in phonology, we felt that limiting the consonant class to the obstruent consonants provided a better conceptual grouping of sounds into classes for these experiments. This grouping also resulted in an equal number of vowel and consonant phonemes. Hereafter, when we refer to "consonants" and "vowels," we mean the grouping of sounds into the classes shown in Table 1.

GROUP	PHONE	TOTAL
consonant	b d g p t k dx q jh ch s sh z zh f th v dh hh hv	20
vowel	iy ih eh ey ae aa aw ay ah ao oy ow uh uw ux er ax ix axr ax-h	20
weakson	l r y w el eng nx m n ng em en	12

Table 1. Classification of Phonemes

In experiment 1, we replaced either the consonant sounds in an utterance with noise, leaving segments in the vowel and weakson groups unaltered, or we replaced the vowel sounds in an utterance with noise, leaving segments in the consonant and weakson groups unaltered. In experiment 2, each utterance consisted of segments from only one of the three groups; for example, if the utterance consisted of vowel sounds, all segments from the consonant and weakson groups were replaced by noise. In experiment 3, we replicated experiment 1, but added four additional conditions by including utterances in which either the consonant or vowel boundaries in the sentences were expanded or reduced by 10 msec before replacing segments with noise. In all three experiments, having the original vowel information available resulted in much better recognition. The size of the effect is dramatic. For example, in experiment 2, when vowels were the only unaltered segments, 56.5% of the words and 21.5% of the sentences were still recognized. When consonants alone were unaltered, only 14.4% of the words and none of the sentences were recognized.

2. EXPERIMENT 1

Experiment 1 was our first attempt to understand the relative contribution of vowels *vs.* consonants to word recogni-

[1]The stimuli used in these experiments are available via ftp from OGI. See http://www.cse.ogi.edu/CSLU for instructions.

0-7803-3192-3/96 $5.00©1996 IEEE

Table 2. Some Statistics of Consonants and Vowels in the 60 Selected Sentences in Experiment I

	VOWELS	CONSONANTS	WEAKSON
#occurrences	756	747	417
duration(ms)	67368	71047	23556

tion. The experiment was designed to (a) remove the spectral information associated with replaced consonants and vowels; (b) retain the the energy and duration profile of the replaced segment and thus minimize changes to the prosodic structure of the utterance; (c) balance the numbers of occurrences of consonants and vowels to the extent possible within each sentence; and (d) balance the total duration of vowel and consonant segments across all sentences.

2.1. Stimulus Preparation

Since consonants and vowels are modeled as being generated by noise and periodic sources, respectively, we sought to eliminate possible bias created by the nature of the substituting sound by using two types of replacement sounds: white noise, and a periodic sound composed of sinusoids with frequencies ranging from 200Hz to 4KHz. Speech sounds were replaced by substituting successive 5 msec intervals of the replaced segment with a replacement signal of the same amplitude.

The manipulations produced five versions of each utterance:

- **CLN**: CLeaN—The original utterance,
- **NCW**: No Consonants (White noise substituted),
- **NCP**: No Consonants (Periodic noise substituted),
- **NVW**: No Vowels (White noise substituted), and
- **NVP**: No Vowels (Periodic noise substituted).

2.2. Data Selection

Sixty sentences were selected from TIMIT database, spoken by 30 male and 30 female speakers from the DR2 (northern) dialect region—one sentence per speaker. Sentences were selected to have the same number of consonants and vowels, and to have the same total duration of consonants and vowels over the set of 60 sentences. Details are given in Table 2.

2.3. Arrangement of Test Lists

We arranged the processed versions of the 60 sentences into 5 lists, each containing 60 sentences. Within each list, there were 12 sentences from each of the 5 categories (CLN, NCW, NCP, NVW, NVP) in random order. No sentences (from the same original source) were repeated within a list, so subjects never heard the same text twice.

2.4. Experimental Procedure

The experiment was performed on a workstation equipped with audio I/O. Subjects listened to the speech via closed ear-cushion headphones at a comfortable volume level set by the subject. A simple graphical user interface was designed to allow subjects to control the presentation of the sentences and to type the words they heard. The subjects could listen to each sentence up to five times at their own pace. After listening to each sentence, subjects were asked to type in as many words as they could understand, or revise what they had written down in the previous presentation of the same sentence.

Table 3. Word/Sentence Correct Rates for Exp I

CATEGORY	MEAN % CORRECT	
	WORD	SENTENCE
CLN	97.8	86.4
NCW	87.4	60.0
NCP	81.9	49.8
NVP	47.9	11.9
NVW	46.6	10.7

Thirty-five high school graduates served as subjects. All of the subjects were native American English speakers with no reported hearing problems. Before beginning the experiment, each subject was given a training session with examples of utterances from each of the five categories.

2.5. Results and Discussion

Subjects' responses were spell checked, and a dynamic programming string-alignment algorithm was used to calculate the word and sentence recognition rates. Word mismatches caused by the inherent ambiguity of English were tolerated. For example, "shellfish" and "shell fish," were treated as equivalent. At the sentence level, only those sentences with all the words exactly matching the original TIMIT texts were considered to be correctly understood.

Results of experiment I are summarized in Table 3. When vowels are available (in addition to liquids, glides and nasals), almost twice as many words are recognized as when consonants are available (in addition to liquids, glides and nasals). Viewed in terms of word error rate, about five times as many errors occur when listeners are presented with consonants and weak sonorants, compared to vowels and weak sonorants. Subjects are able to recover all of the words in over half of the sentences when vowels are available, and in only about 11% of the sentences when consonants are available.

Analysis of variance showed a significant effect of segment type (consonants vs. vowels; $p < 0.01$) and no effect of the type of substituting noise for the segment. In the remaining experiments, we used white noise as the replacement sound.

3. EXPERIMENT 2

One possible explanation of the results in experiment 1 was that leaving the weak sonorants in place somehow was more beneficial to the vowels than to the consonants. Also, it was unclear to what extent the weakson group contributed to recognition. Experiment 2 was performed to assess the relative contribution of consonants, vowels and weak sonorants to word recognition, when the spectral information from these categories is the *only* information available to the listener.

3.1. Stimulus Preparation

In experiment 2, rather than omitting one group of segments, we preserved one group, and replaced segments in

Table 4. Word/Sentence Performance for Exp 2

CATEGORY	MEAN % CORRECT	
	WORD	SENTENCE
CLN	94.0	74.4
C	14.4	0.0
V	56.5	21.5
W	3.1	0.0

the two remaining categories with white noise. This resulted in 60 sentences divided among four groups:

- **CLN**: CleaN—The original utterance.
- **C**: Consonants only (vowels and weakson replaced by noise).
- **V**: Vowels only (consonants and weakson replaced by noise).
- **W**: Weakson only (vowels and consonants replaced by noise).

We arranged the processed versions of the sixty sentences into four lists, each containing 60 sentences. Within each list, there were 15 sentences from each of the four mentioned categories (CLN, C, V, W) in random order. No sentences (from the same original source) were repeated within a list, so subjects were never presented the same utterance twice.

3.2. Experimental Procedures

The experimental procedures were identical to experiment 1. Thirteen high school graduates who were native speakers of American English, reported no hearing problems, and who had not participated in experiment one, served as subjects. The four test lists were distributed (almost) equally among the subjects.

3.3. Results and Discussion

Table 4 shows that listeners identify 56.5% of the words in TIMIT sentences when vowel information is available, and all other segments are replaced by noise, compared to 14.4% for consonants and 3.1% for weak sonorants. Moreover, vowel information alone was sufficient to recognize all of the words in 21.5% of the sentences, whereas no sentences could be recognized completely using only consonants or weak sonorants. T-tests on the word and sentence scores showed that all of the observed differences were statistically significant (p=.001).

4. EXPERIMENT 3

In experiment 3 we begin to investigate the basis for the large effect observed in experiments 1 and 2. A possible explanation for the greater importance of vowels to word recognition is that coarticulatory information in vowels provides enough information about adjacent consonants to allow listeners to recover the intended words. It is well known that formant transitions at vowel onsets and offsets play a key role in the perception of adjacent consonants. Perhaps coarticulatory information at vowel boundaries provide more information about consonants, than *vice versa*.

To test this hypothesis, we replicated experiment 1, but added 4 new experimental conditions, in which we either expanded or reduced vowel segments or consonants segments by moving segment boundaries by 10 msec as shown in Figure 1. If formant transitions play a key role in the observed effect, expanding consonant boundaries by 10 msec in each direction should produce the greatest improvement in word recognition performance, since information about the vowel is now included in the consonant, whereas expanding vowel boundaries should provide relatively less new information.

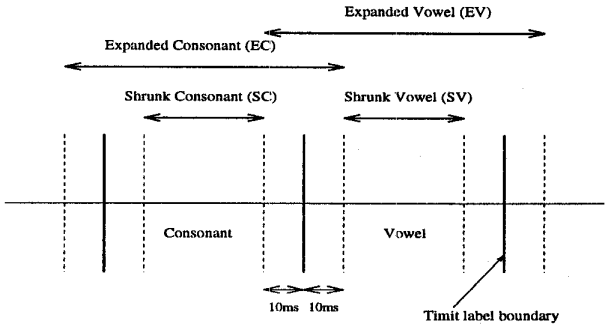

Figure 1. An example of Consonant/Vowel boundary modification

4.1. Stimulus Preparation

Starting with the time-aligned phonetic transcription of a TIMIT sentence, the unaltered segments in a sentence were expanded by 20 msec (10 msec from each boundary), or reduced by 20 msec (10 msec from each boundary). The inclusion of expanded and reduced segments produced 7 versions for each utterance:

- **CLN**: CLeaN—the original utterance,
- **EV**: Expanded Vowels (shrunk consonants substituted with white noise), and
- **V**: Vowels (consonants substituted with white noise),
- **SV**: Shrunk Vowels (expanded consonants substituted with white noise).
- **EC**: Expanded Consonants (shrunk vowels substituted with white noise),
- **C**: Consonants (vowels substituted with white noise),
- **SC**: Shrunk Consonants (expanded vowels substituted with white noise),

4.2. Data Selection

Eighty-four sentences were selected from the TIMIT database. The selection criteria were the same as we used for selecting the 60 sentences used in Experiment 1.

After processing, each original sentence yields 7 utterances, including the original. As in the first experiment, the processed utterances were assigned to seven lists, each containing 84 sentences. Within each list, there were 12 sentences from each of the seven categories, in random order. No sentences (from the same original source) were repeated within a list.

4.3. Experimental Procedure

Fourteen subjects, different from those who participated in experiments 1 and 2, were selected according to the same criteria. The experimental procedure was identical to the previous experiments, except in one respect. Due to the increase in the number of testing sentences, the experimental

Table 5. Word/Sentence Performance for Exp 3

CATEGORY	MEAN % CORRECT	
	WORD	SENTENCE
CLN	96.3	76.8
EV	87.2	59.5
V	84.2	48.2
SV	79.8	41.7
EC	53.1	14.9
C	41.8	10.7
SC	26.7	3.0

session for each subject was divided into two parts, (separated by at least 30 min) to prevent possible fatigue.

4.4. Results and Discussion

Results of experiment 3 are summarized in Table 5. Adding 10msec of the adjacent segment results in about a 20% reduction in word recognition error for both consonants and vowels. Deleting 10msec of the retained segment causes about a 25% increase in word error for both consonants and vowels.

Statistical analysis shows that the performance differences between SC *vs* C, and C *vs* EC are significant at 0.01 level, while the performance differences between SV *vs* V, and V *vs* EV are not significant at 0.05 level but at 0.1 level.

5. SUMMARY AND DISCUSSION

We tested the relative importance of vowels and obstruent consonants to recognition of fluent speech by replacing one or the other group with noise in perceptual experiments. In experiment 1, liquids, glides and nasals ("weak sonorants") were left unaltered and either vowels or the remaining consonants were replaced by noise of the same duration and energy as the replaced speech. Both white noise and periodic noise were used; no significant difference between the two was found nor was there an interaction with the type of replaced segment. When only vowels and the weak sonorants were present, subjects couldn't recognize 13% of the words. When only consonants and weak sonorants were present, subjects couldn't recognize 52% of the words.

In experiment 2, the weak sonorants were replaced by noise as well as either vowels or consonants. With vowels alone, subjects couldn't recognize 44% of the words. With consonants alone, subjects couldn't recognize 86% of the words. The effect of vowels being superior to consonants remains with or without the presence of weak sonorants. Experiment 2 also showed that weak sonorants alone (vowels and consonants both replaced by noise) results in very low recognition—only 3% of the words were recognized. Even though this group is inadequate on its own, combined with vowels *or* consonants they improve recognition a great deal (comparing recognition in experiment 1 with that in experiment 2).

In experiment 3, we tested the hypothesis that the edges of vowels contain more information about the neighboring consonants than the edges of consonants contain about neighboring vowels. Segment boundaries were expanded or contracted 10 msec before replacing the segments with noise. If the hypothesis were true, we would expect greater damage to performance when the edges of vowels are removed than when the edges of consonants are removed and more improvement by expanding consonants into vowels than by expanding vowels into consonants. The results were inconclusive. The magnitude of the effect (as it affected error rate) was approximately the same for both vowels and consonants. The improvement from expanding consonants tested significant whereas expanding vowels did not, which supports the hypothesis. The harm from shrinking consonants tested significant whereas shrinking vowels did not, which fails to support the hypothesis. A possible explanation is that, since formant transitions are typically longer than 10 msec they convey information about adjacent consonants even when 10 msec of the transition is removed.

There is a clear, unambiguous and overwhelming conclusion: vowels are more important for recognition than the obstruent consonants, despite the fact that they are equally represented in the test sentences. We investigated an acoustic basis for this effect—that vowels contain more coarticulatory information that can be used to recover the missing consonants than vice versa. There are other possible explanations. For example, the amplitude envelope information in the replacement sounds may provide more information about consonants than vowels. It may also be the case that vowels simply convey more information about words than consonants from an information theoretic point of view. Further research is needed to resolve the basis of this effect

REFERENCES

[1] L.F. Lamel, R.H. Kassel, and S. Seneff,"Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus",*Proc. DARPA Speech Recognition Workshop* , Report No. SAIC86/1546, pp.100-109, February 1986.