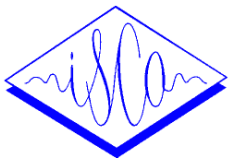

ESCA Workshop on
Audio-Visual Speech Processing
Rhodes, Greece
September 26-27, 2997

EFFECTS OF PHONETIC VARIATION AND THE STRUCTURE OF THE LEXICON ON THE UNIQUENESS OF WORDS

E. T. Auer, Jr., L. E. Bernstein, R. S. Waldstein, and P.E. Tucker
Spoken Language Processes Laboratory
House Ear Institute
2100 West Third Street
Los Angeles, California 90057

ABSTRACT

Relatively little is known about the *optical phonetic* speech characteristics to which speechreaders are attuned. However, it is known that phonetic context can affect visual confusability of phonemes. In Study 1, behavioral experiments were performed to examine in detail effects of context-sensitive phonetic variation on the visual confusability of consonants and vowels. In Study 2, computational experiments were performed to assess the importance of patterns of context-sensitive visual confusability on the uniqueness of words in the language. Results from Study 1 further support the conclusion that phonetic context influences phoneme confusability. The computational experiments in Study 2 provide evidence that the distribution of words in English substantially preserves lexical uniqueness even when phonetic variability is taken into account.

1. INTRODUCTION

Visible speech generally affords less phonetic information than acoustic speech. As a result, the speechreader may only perceive a reduced set of phonemic distinctions. Previously, Auer and Bernstein (in press) investigated the relationship between visually perceivable phonemic distinctions and the predicted uniqueness of speechread words in English. Auer and Bernstein demonstrated that the distribution of words in English substantially preserves lexical uniqueness, and that estimates of lexical uniqueness were sensitive to small changes in the number of available phonemic distinctions. For example, the loss of the phonemic distinctions among /b/, /p/, and /m/ results in a loss of lexical uniqueness for the words "bat", "pat", and "mat." However, the word "bought" remains unique, because "pought" and "mought" are not words in English.

In the Auer and Bernstein analyses, phonetic similarity was estimated based on identification of phonemes in monosyllabic nonsense syllables. However, coarticulation effects arising from variation in surrounding phonetic contexts have been demonstrated to alter phoneme identification by speechreaders (Benguerel & Pichora-Fuller, 1982; Jackson, 1988; Owens & Blazek, 1985). Thus, the set of nonsense syllable identifications employed by Auer and Bernstein could have lead to either under- or over-estimation of the phonetic information available to the speechreader and therefore inaccurate estimates of lexical uniqueness.

The current paper is a first report on two studies that address the possibility that context-sensitive variation is important for the uniqueness of speechread words. In Study 1, phoneme identification experiments with stimuli controlled for phonetic context were performed. In Study 2, computational experiments were conducted to evaluate the importance of the patterns of context-sensitive visual-phonemic confusability observed in Study 1.

2. STUDY 1

This study was designed to estimate effects of phonetic context on the visibility of phonemes in skilled deaf speechreaders.

2.1. Methods

2.1.1. Participants

Study participants were 18 adults, aged 18-30 years, with severe or profound congenital hearing losses [3-frequency (500, 1000, and 2000 Hz) pure-tone averages ≥ 80 dB HL bilaterally]. All participants were required to meet three criteria: 1) education either in oral programs for deaf children or in mainstream settings, where English was the language of instruction, for a minimum of 8 years; 2) 20/30 or better vision in both eyes; and 3) average or better lipreading ability. Five individuals participated in the initial consonant and in the vowel studies; and four participated in the medial consonant and in the final consonant studies. No individual participated in more than one study; all had earned at least a high school degree; and none had any known learning disabilities.

2.1.2. Stimuli

Initial consonants: two tokens each of the 27 consonants or consonant clusters /b p m f v ð θ tʃ dʒ ʃ w r d h g k l n s z t j pr st tr gr kr/ in the four contexts, C-/adəd/, C-/idəd/, C-/udəd/, and C-/ədad/ (216 tokens total).

Medial consonants: two tokens each of the 27 consonants or consonant clusters /b p m f v ð θ tʃ dʒ ʃ ʒ w d h g k l n s z t j nd ns nt kt st/ in the six contexts, /də/-C-/ad/, /də/-C-/id/, /də/-C-/ud/, /da/-C-/əd/, /di/-C-/əd/, and /du/-C-/əd/ (324 tokens total).

Final consonants: two tokens each of the 24 consonants or consonant clusters /b p m f v ð θ tʃ dʒ ʃ ʒ d g k l n s z t lz nz nd nt st/ in the four contexts /dəda/-C, /dədi/-C, /dədu/-

C, /dadə/-C, and two tokens of /ŋ/ in the two contexts, /dəda/-C and /dadə/-C (196 tokens total).

Vowels: two tokens each of the vowels /i ɪ ɛ æ a ɔ ʌ ʊ u/, r-colored vowels / ɜ˞ ɪr ʊr ar er/, and diphthongs /eɪ oʊ aʊ aɪ ɔɪ/, in the four contexts, /m/-V-/m/, /n/-V-/n/, /p/-V-/p/, and /t/-V-/t/ (152 tokens total).

Consonant clusters comprised the five most frequently occurring clusters in a given syllabic position. A female adult, native speaker of English, served as the talker for the studies. She was professionally videotaped in color against a neutral background with her head filling the screen. The stimuli were dubbed onto optical videodisks for presentation.

2.1.3. Procedure

Each study employed the same protocol. Participants were tested individually in a quiet room. They were presented with their assigned randomized stimulus set for a total of 10 times, each time during a separate test session. Participants watched the stimuli on a 14-inch color monitor. They were instructed to identify the target phoneme in the spoken nonsense word, and to indicate their choice by pressing the appropriately labeled key on the keyboard in front of them. Feedback was provided on all trials. For each study, a list of the labels and a sample word featuring the corresponding phoneme was in full view during testing.

2.2. Results and Discussion

Tables 1-4 summarize the preliminary analyses of the stimulus-response confusion matrices. The columns in each table represent percent correct for the matrix, number of phonemic equivalence classes (P.E.C.; see definition of P.E.C. in Study 2), and phonetic environments, and the asterisks indicate significant contrasts based on log-linear analyses (see below).

Context	%C	P.E.C.	/a/	/i/	/ə/	/u/
/a/	43	11				*
/i/	36	4				*
/ə/	31	5				*
/u/	35	6	*	*	*	

Table 1. INITIAL consonants. (2700 responses collected in each context, total = 10800)

Context	%C	P.E.C.	əCa	aCə	əCi	iCə
əCa	25	6				
aCə	23	6				
əCi	23	5				
iCə	24	6				
əCu	21	5	*	*	*	*
uCə	20	4	*	*	*	*

Table 2. MEDIAL consonants. (2160 responses collected in each context, total = 12960.)

Context	%C	P.E.C.	/a/	/i/	/ə/	/u/
/a/	42	7				*
/i/	36	6			*	*
/ə/	29	5		*		*
/u/	33	4	*	*	*	

Table 3. FINAL consonants. (1920 responses collected in the /i/ and /u/ context and 2000 responses collected in the /a/ and /ə/ context, total=7840.)

Context	%C	P.E.C.	m	p	n	t
mVm	75	12				
pVp	76	11				
nVn	75	12				
tVt	74	10				

Table 4. VOWELS. (1900 responses were collected in each context, total = 7600.)

The tables show that vowel identification was more accurate than consonant identification. Identification accuracy in initial and final position was roughly equivalent, and was poorest for consonants in medial position.

Vowel identification accuracy did not vary as a function of the surrounding consonantal context. However, consonant identification accuracy varied as a function of vowel context. Consistent with previous studies, consonant identification accuracy was worst in the /u/ environment, a likely result of lip rounding for /u/.

Log-linear analyses (Bishop et al., 1980) were used to evaluate whether stimulus-response patterns varied across contexts. Analyses indicated no differences in the response patterns of the VOWEL stimulus set. In the INITIAL, MEDIAL, and FINAL consonant stimulus sets, response patterns for consonants produced in the /u/ environment differed from response patterns collected in all the other environments. In addition, in the FINAL stimulus set, response patterns from the /i/ and /ə/ environments differed from each other.

3. STUDY 2

Within visual speech perception research, phonetic context has been viewed as detrimental to accurate perception (e.g. Owens & Blazek, 1985). Church (1987) has argued (for acoustic speech) that effects of coarticulation are informative for word recognition. One way to estimate the effect of phonetic context on intelligibility is to computationally evaluate its effect on the uniqueness of words in the lexicon. In Study 2, word uniqueness was modeled under a variety of levels of phonetic-context sensitivity.

3.1 Methods

Computational lexical modeling techniques (Altmann, 1990; Auer & Bernstein, in press; Carter, 1987) were applied as follows: First, a phonemically transcribed machine-readable **lexical database** was selected to serve as a representative sample of the words in the language.

Second, **transcription rules** were defined in the form of symbol substitutions for all phonemes in phonemic equivalence classes. A **phonemic equivalence class** comprised the set of phonemes or clusters modeled as mutually confused using the behavioral data from Study 1 (see Table 5). Third, the lexical database was then transcribed by applying the transcription rules. **Lexical equivalence classes** were formed by collapsing across identically transcribed words. Finally, metrics were computed to compare the distribution of patterns in the newly transcribed lexicon against the distribution of patterns in the original lexicon.

3.1.1. Lexical Database

The method described above was applied to the PhLex database (Seitz et al., 1995), which comprises the 20,000 most frequent words in Kucera and Francis (1967) and the 12,118 words in Nusbaum et al. (1984). PhLex's entries have transcriptions with stress and syllabification markers and estimates of frequency of usage. When word frequency information was not available for an entry, frequency was set to 1. All frequencies were log-transformed (base 10).

3.1.2. Transcription Rules

Sets of transcription rules were developed using the stimulus-response confusion matrices obtained in Study 1. Separate Phi-square analyses were performed to generate estimates of similarity from the count data in each matrix. These estimates were then submitted to separate hierarchical cluster analyses using the average linkage between groups method. The agglomeration schedule generated by the cluster analysis was then used to combine segments into phonemic equivalence classes until all phonemic equivalence classes in a particular matrix included 75 percent or more within-cluster responses corresponding to the typical viseme level of confusability (see Bernstein et al., these proceedings for arguments against this level).

Context	Phonemic Equivalence Classes
VOWELS	{i,ɪ} {ɛ,æ} {a, ʌ, ə} {ɔ,ar,ɔr} {ɝ,ʊ,ʊr} {u} {ɪr} {er} {eɪ} {oʊ} {aʊ} {aɪ} {ɔɪ}
INITIAL /a/	{b,p,m} {pr} {f,v} {ð,θ} {ʃ,tʃ,ʒ,dʒ} {w} {d,t,s,z,n,k,ŋ,g,j st} {h} {r,gr,kr} {l} {tr}
MEDIAL /aCə/+/əCa/	{b,p,m} {f,v} {ð,θ} {ʃ,tʃ,ʒ,dʒ} {w} {d,t,s,z,n,k,ŋ,g,j,h,l,kt,st,ns,nt,nd}
FINAL /a/	{b,p,m} {f,v} {ð,θ} {ʃ,tʃ,ʒ,dʒ} {d,t,s,z,n,st,nz,nt,nd} {g,k,ŋ,l} {lz}
INITIAL /u /	{b,p,m,pr} {f,v} {ð,θ} {w,r,gr} {l} {ʃ,tʃ,ʒ,dʒ,d,t,s,z,n,k,h,ŋ,g,j,st,tr,kr}

Table 5. Phonemic equivalence classes for vowels and consonants in different contexts used for generating transcription rules.

3.1.3. Application of Transcription Rules

Auer and Bernstein (in press) estimated, using phoneme identifications in C/a/ environments, that more than 54 percent of words should be visually unique to speechreaders. Study 1 implies that such estimates might require taking into account phonetic environment of phonemes. Therefore, in Study 2, as a preliminary investigation of phonetic environment effects, five sets of transcription rules (Table 6) were applied to the PhLex Database. For sets 1, 2, and 5, transcription rules were based on consonant-initial identifications: consonants in all positions were transcribed as if they were as intelligible as consonants in initial position. Sets 1 and 2 were based on identification of consonants in a single vowel context, /a/ and /u/ respectively. Set 5 used identification of consonants in both the /a/ and /u/ context: the transcription rules from consonants in /u/ context were only applied to syllables containing the vowels /u,ɝ,ʊ,ʊr/, and transcription rules from consonants in /a/ context were applied to all other consonants. Set 3 was based on identifications of consonants in MEDIAL position in the /a/ vowel context. Set 4 was based on identification of consonants in INITIAL, MEDIAL, and FINAL position in the /a/ vowel context. INITIAL position was defined as consonants prior to the first vowel in the word. MEDIAL position was defined as consonants occurring after the first vowel and before the last vowel in the word, and FINAL position was defined as consonants occurring after the last vowel in the word. Sets 1-5 used the same set of vowel phonemic equivalence classes.

Phonemic equivalence classes that contained both single consonants and consonant clusters were transcribed using single symbol substitution. Phonemic equivalence classes that contained consonant clusters alone were transcribed to preserve their two segment length, and no subsequent transcription rules were applied to the clusters' component segments. Two words were considered equivalent only when their phonemic, stress, and syllabification patterns were identical (e.g., the noun "convert" and the verb "convert" were not considered equivalent), thus assuming accurate perception of lexical stress and syllabification.

3.1.4. Quantitative Analysis

Two metrics were computed to quantitatively analyze the distributions of patterns in the transcribed lexicon (Auer & Bernstein, in press; Carter, 1987). Frequency-weighted percent words unique was the sum of the frequencies of occurrence for unique words in the transcribed lexicon divided by the sum of frequencies of occurrence of words in the original lexicon. This metric estimated the extent to which unique words are encountered in everyday language.

Frequency-weighted expected class size was computed as

$$ECS = \sum_{a=1}^{n_E} I_a \frac{F_a}{F_L}, \qquad (1)$$

where n_E is the total number of lexical equivalence classes, I_a is the number of words in equivalence class a, F_a is the sum of frequencies of occurrence of words in

equivalence class a, and F_L is the sum of the frequencies of occurrence of words in the lexicon. The frequency-weighted metric estimated the average size of the equivalence classes encountered in everyday language.

3.2. Results and Discussion

Table 6 shows results. Auer and Bernstein's (in press) results were thought to potentially over-estimate lexical uniqueness due to the reasonably high intelligibility of consonants in C/a/ nonsense syllables (similar to Set 1). In the current study, Sets 2 and 3 modeled poor consonantal intelligibility. Set 3 was used to model consonant intelligibility in a context similar to running speech, in which most consonants occur medially. Sets 4 and 5 more closely modeled variations in consonantal intelligibility as a function of phonetic environment. Although differences in both measures were obtained across transcription sets, all transcription rule sets resulted in substantially preserved lexical uniqueness.

Transcription Rule Sets	Percent Unique Words	Expected Class Size
Set 1: INITIAL /a/	57.7	4.25
Set 2: INITIAL /u/	49.6	8.16
Set 3: MEDIAL /a/	49.4	8.50
Set 4: INITIAL, MEDIAL,and FINAL /a/	54.7	4.11
Set 5: INITIAL/a/ and INITIAL /u/	57.1	4.37

Table 6. Frequency-weighted percent unique words and expected class size as a function of transcription rule set.

4. CONCLUSIONS

Study 1 supports the conclusion that accuracy of consonant identification varies both as a function of vowel context and position. Among stimulus-response patterns to consonants, ones in the context of vowel /u/ differed from all others. The computational experiments in Study 2 support the conclusion that the distribution of phoneme patterns in English substantially preserves lexical uniqueness, even when variability due to phonetic context is taken into account.

This research was supported by a grant from the US National Institutes of Health (DC02107).

5. REFERENCES

Altmann, G. T. M. (1990), "Lexical statistics and cognitive models of speech processing," in *Cognitive Models of Speech Processing*, G.T.M. Altmann, Ed., MIT Press, Cambridge, pp. 211-235.

Auer, Jr., E. T. & Bernstein, L. E. (in press), *"Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness"*

Bengeurel, A. P. & Pichora-Fuller, M. K. (1982), *"Coarticulation effects in lipreading,"* J. Speech Hear. Res. 25, pp. 600-607.

Bernstein, L. E. , Iverson, P. I., & Auer, Jr., E. T. (1997), "Elucidating the complex relationships between phonetic perception and word recognition in audiovisual speech perception," *Proc. AVSP'97*, Rhodes (Greece).

Bishop, Y. M. M., Fienberg, S. E., & Holland, P.W. (1980), *'Discrete Multivariate Analysis, Theory and Practice,'* MIT Press, Cambridge, MA.

Carter, D. (1987), *"An information theoretic analysis of phonetic dictionary analysis,"* Comput. Speech Lang. 2, pp. 1-11.

Church, K.W. (1987), *'Phonological Parsing in Speech Recognition,'* Kluwer Academic Publishers, Boston, MA.

Jackson, P. L. (1988), "The theoretical minimal unit for visual speech perception: Visemes and coarticulation," in *New Reflections on Speechreading*, C. DeFillipo & D. Sims, Eds., A.G. Bell Association for the Deaf, Washington, DC, pp. 99-115.

Kucera, H. & Francis, W. (1967), *"Computational Analysis of Present-Day American English,"* Brown University Press, Providence.

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984), "Sizing up the Hoosier Mental Lexicon: Measuring the familiarity of 20,000 words," *Research on Spoken Language Processing PR-10*, Indiana University, Bloomington, IN.

Owens, E. & Blazek, B. (1985), *"Visemes observed by hearing-impaired and normal-hearing adult viewers,"* J. Speech Hear. Res., 28, pp. 381-393.

Seitz, P. F., Bernstein, L. E., and Auer, Jr., E. T. (1995), *PhLex (Phonologically Transformable Lexicon), A 35,000-word pronouncing American English lexicon on structural principles, with accompanying phonological rules and word frequencies*, Gallaudet Research Institute, Washington, DC.