LIPREADING SUPPLEMENTED BY VOICE FUNDAMENTAL FREQUENCY: TO WHAT EXTENT DOES THE ADDITION OF VOICING INCREASE LEXICAL UNIQUENESS FOR THE LIPREADER?

Edward T. Auer, Jr. and Lynne E. Bernstein

Spoken Language Processes Laboratory
House Ear Institute
2100 West Third Street
Los Angeles, California 90057

ABSTRACT

Lipreading in combination with an acoustic indication of voice fundamental frequency (F0) has been shown to greatly enhance word recognition accuracy with sentence stimuli [1]. A possible explanation for this effect is that F0 delivers information for consonantal voicing. In Experiment 1, we showed with a computational model how voicing information affects the uniqueness of lipread words in a large phonemically transcribed machine-readable lexicon. In Experiment 2, the same computational methods were used to simulate the results obtained by McGrath and Summerfield [2] for lipreading with and without acoustic F0. The model failed to account in full for the behaviorally observed enhancements. It is suggested that lexical biasing in word recognition can account for the difference between the model and the behavioral results. (This work was supported by NIH Grant DC-00695.)

1. INTRODUCTION

Even under optimal viewing conditions, not all phonetic information is visible to the lipreader. As a result, the information needed to perceive some phonemic distinctions is not available. For example, lipreaders may not perceive any distinctions among productions of the consonants /b/, /p/, and /m/. The loss of phonemic distinctions results in reduced uniqueness for words in the lexicon [3,4]. Thus, understanding spoken language is difficult for many deaf individuals. In order to enhance lipreading by deaf individuals, investigators have sought signals that can be transduced by an impaired auditory system or by an alternate sensory system such as touch.

One such signal is voice fundamental frequency (F0). F0 is generated at the glottis, which is invisible to the lipreader. Several experiments have been reported in which simple *acoustic* stimuli composed of pulses generated as a function of F0 were presented to enhance lipreading. In these experiments, adults with normal hearing improved as much as 40 percentage points over lipreading alone when they lipread with the F0 supplement [1,2].

The observed enhancement is typically attributed to the fact that F0 characteristics contribute to perception at several different linguistic levels, including consonantal voicing distinctions [5], lexical stress (e.g., CONvert versus conVERT), sentential stress [6], word boundaries [7] and syntactic information [6]. However, estimates of the contribution made by the different characteristics associated with F0 to the overall enhancement effect have not been obtained. Waldstein and Boothroyd [8] have suggested that as much as one half of the observed enhancement may be due to the information conveyed about the presence of consonantal voicing. The current computational experiments examined the contribution of consonantal voicing to the uniqueness of words in the lexicon.

1.1 Sources of Voicing in Lipreading

Because laryngeal vibrations are invisible, consonantal voicing is frequently hypothesized to be completely absent from the information available to the lipreader. Although this assertion may be true for lipread consonant-vowel nonsense syllables, it is not true for lipread words or sentences [9].

One source of consonantal voicing information is the preceding vowel duration for post-vocalic consonants [10]. Vowel durations are longer for voiced final consonants than for voiceless consonants. Durational cues are potentially available to the lipreader and are likely responsible for the partial visibility of final consonant voicing reported by Hnath-Chisolm and Kishon-Rabin [11]. Another source of consonantal voicing information is the distribution of phoneme patterns in the words of the language. For example, /b/ is distinguished from /p/ or /m/ in the English word “bought,” because “pought” and “mought” are not words. Thus, the voicing distinction is available by virtue of the lexicon’s structure. Of course, the voicing distinction is not disambiguated via the lexicon’s structure for all words (e.g. “bat”) [4].

2. EXPERIMENT 1

The goal of Experiment 1 was to model effects on the structure of the lexicon brought about when visible speech is enhanced with consonantal voicing information. Computational lexical modeling techniques [4,12,13] were applied to obtain frequency-weighted estimates of word uniqueness for lipreading alone (LA) and lipreading with voicing information (L+V).

2.1 Methods

Lexical modeling was 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. Along with a phonemic transcription, each word in the database had an associated estimate of its frequency of occurrence in the language. Second, **transcription rules** were defined on the basis of measures of phonetic similarity. The transcription rules were in the form of single symbol substitutions for all phonemes in phonemic equivalence classes. A **phonemic equivalence class** comprised the set of phonemes rendered equivalent by the loss of phonetic distinctiveness. (For example, when /b/, /p/, and /m/ are phonetically similar, a transcription rule is defined to transcribe each occurrence of /b/, /p/, and /m/ into one symbol representing the equivalence class.) Third, the lexical database was then transcribed by applying the transcription rules. **Lexical equivalence classes** were formed by collapsing across identically transcribed words. (For example, under the phoneme equivalence class definition given above, "pat" and "bat" would both fall into the same lexical equivalence class.) Finally, metrics were computed to compare the distribution of patterns in the newly transcribed lexicon with the distribution of patterns in the original lexicon.

Lexical Database. The method described above was applied to the PhLex database [14], which comprises the 20,000 most frequent words in [15] and the 12,118 words in [16]. All of PhLex's entries have transcriptions that include 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).

Transcription Rules. Sets of transcription rules were developed using estimates of visual phonetic similarity obtained from separate behaviorally obtained consonant and vowel confusion matrices [17]. These estimates were submitted to separate hierarchical cluster analyses using the average linkage between groups method. Because perceptual data were not available for /ə j ŋ/, theoretical estimates of similarity were employed. Vowels and consonants were assumed to be maximally dissimilar, except for the consonant /j/ which was included in the vowel confusion matrix. (See Table 1.) The transcription rules applied to 17 vowels, and 23 consonants.

Table 1 lists the sets of phonemic equivalence classes that were used for the transcription rules for the LA condition. The table shows that the number of equivalence classes increased at the same rate for consonants and vowels, and that the increases followed the hierarchical clustering results for between 2 and 19 clusters. The range between 10 and 19 clusters best approximates the phoneme equivalence classes estimated for lipreaders [4].

A second group of transcription rule sets was generated for the L+V condition. This was accomplished by modifying each equivalence class such that voiceless consonants never appeared in the same equivalence class with a voiced or nasal consonant. For example, the phonemic equivalence class {d,t,s,z} was separated into two new equivalence classes, {t,s} and {d,z}.

Number of Phonemic Equivalence Classes	Phonemic Equivalence Classes
19	{u,ʊ,ər} {o,aʊ} {ɪ,i} {e,ɛ} {æ} {ɔɪ} {ɔ} {aɪ,ə,ɑ,ʌ,j}{b,p,m} {f,v} {l} {n,k} {ŋ,g} {h} {d} {t,s,z} {w,r} {ð,θ} {ʃ,tʃ,ʒ,dʒ}
12	{u,ʊ,ər} {o,aʊ} {ɪ,i,e,ɛ,æ} {ɔɪ} {ɔ,aɪ,ə,ɑ,ʌ,j} {b,p,m} {f,v} {l,n,k,ŋ,g,h} {d,t,s,z} {w,r} {ð,θ} {ʃ,tʃ,ʒ,dʒ}
10	{u,ʊ,ər} {o,aʊ} {ɪ,i,e,ɛ,æ} {ɔɪ,ɔ,aɪ,ə,ɑ,ʌ,j}{b,p,m} {f,v} {l,n,k,ŋ,g,h,d,t,s,z} {w,r} {ð,θ} {ʃ,tʃ,ʒ,dʒ}
2	{u,ʊ,ər,o,aʊ,ɪ,i,e,ɛ,æ,ɔɪ,ɔ,aɪ,ə,ɑ,ʌ,j} {b,p,m,f,v,l,n,k,ŋ,g,h,d,t,s,z,w,r,ð,θ,ʃ,tʃ,ʒ,dʒ}

Table 1. Equivalence classes comprising transcription rules.

Application of Transcription Rules. Transcription rule sets for both LA and L+V were applied to the PhLex database. Two words were considered equivalent only when their phonemic, and stress and syllabification patterns were identical. For example, the noun "convert" and the verb "convert" were not considered equivalent. Thus, these analyses assumed accurate perception of lexical stress and syllabification.

Quantitative Analysis. Two commonly employed metrics were computed to quantitatively analyze the distributions of patterns in the transcribed lexicon [12,13]. Frequency-weighted percent words unique was computed as

$$\%WU = \frac{F_U}{F_L} \times 100 , \qquad (1)$$

where F_U is the sum of the frequencies of occurrence for unique words in the transcribed lexicon, and F_L is the sum of frequencies of occurrence of words in the original lexicon. The frequency-weighted metric estimates the extent to which unique words are encountered in everyday language.

Frequency-weighted expected class size is computed as

$$ECS = \sum_{a=1}^{n_E} I_a \frac{F_a}{F_L} , \qquad (2)$$

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 estimates the average size of the equivalence classes encountered in everyday language.

2.2 Results and Discussion

Table 2 shows that consonantal voicing substantially increases the percent unique words for every set of transcription rules. The largest enhancement was 15 percentage points, when the number of phonemic equivalence classes was 10 for LA and 15 for L+V. We estimated that 10 equivalence classes is typical of relatively inaccurate hearing lipreaders. The table also shows that many words are not unique under the L+V condition, although a substantial reduction in the frequency-weighted expected class size occurs with consonantal voicing (L+V).

Number of Phonemic Equivalence Classes		Percent Unique Words		Expected Class Size	
LA	L+V	LA	L+V	LA	L+V
2	3	7	18	422.6	86.6
10	15	43	58	14.1	4.3
12	18	54	62	5.1	2.2
19	25	76	85	1.6	1.2

Table 2. Percent unique words and expected class size as a function of LA versus L+V and number of phonemic equivalence classes.

3. EXPERIMENT 2

Experiment 2 was conducted to compare modeled LA and L+V with empirically obtained results from McGrath and Summerfield's [2] Experiment 1. In their experiment, the number of keywords correct in sentences was measured for LA and L+V. In analyzing their data, McGrath and Summerfield split their subjects into three groups based on lipreading ability (poor, average, and good). They found that the magnitude of enhancement increased as a function of lipreading ability. Column 4 of Table 3 gives the percent keywords correct for poor, average, and good lipreaders in their study (see Figure 1 in [2]) in LA and L+V conditions.

3.1 Methods

Word set. McGrath and Summerfield employed the Bamford-Kowal-Bench (BKB) standard sentence lists [18]. Only the BKB keywords were analyzed here, as was the case in [2]. Of the 1,050 keywords, five were eliminated from the analysis, because they did not exist in any form in the PhLex database. Of the remaining 1,045 words, morphological changes were required on 12 words (8 singularizations, 1 pluralization, 4 verb tense changes) in order to find appropriate entries in PhLex. Each keyword token was counted. Thus, if a word occurred in several sentences, it contributed proportionally to the results.

Procedure. Three different sets of phonemic equivalence classes were selected to simulate the three levels of lipreaders in McGrath and Summerfield [2]. Ten phonemic LA equivalence classes (with corresponding 14 L+V equivalence classes) were used to model poor lipreaders' performance; twelve LA phonemic equivalence classes (18 L+V) were used to estimate average lipreaders' performance; and nineteen LA phonemic equivalence classes (25 L+V) were used to estimate good lipreaders' performance. (See Table 1 for the 10, 12, and 19 LA equivalence classes.) These six rule sets were applied to the extracted BKB keywords and to the entire PhLex database.

Quantitative Analysis. A BKB words was counted as recognized if its transcribed form was unique in the corresponding transcription of PhLex. Percent correct was obtained by dividing the total number of transcribed unique words by the total number of words in the BKB keyword list.

Average equivalence class size for BKB words was computed under the assumption that subjects selected their response words from their total lexicons. Thus, the average equivalence class size was computed by summing those equivalence class sizes for the classes that contained BKB words in the transcribed PhLex and dividing by the total number of BKB words.

3.2 Results

Number of Phonemic Equivalence Classes		Modeled Percent Correct		Average Equivalence Class Size		Percent Correct McGrath & Summerfield	
LA	L+V	LA	L+V	LA	L+V	LA	L+V
10	15	12	18	34.9	10.2	Poor 9	11
12	18	18	26	12.1	4.5	Avg. 21	38
19	25	45	60	2.6	1.6	Good 42	69

Table 3. Modeled percent words correct, average equivalence class size, and percent words correct [2], as a function of LA versus L+V and number of phoneme equivalence classes.

The results of Experiment 2 are shown in Table 3. The table can be used to obtain the modeled enhancement (L+V minus LA) for BKB words, which is approximately half that of the enhancement reported by [2] for average and good lipreaders. For example, the modeled enhancement for good lipreaders was 15 percentage points (60 minus 45), but the behaviorally obtained enhancement was 27 percentage points (69 minus 42). On the other hand, McGrath and Summerfield's poor lipreaders scarcely benefited from F0, whereas in the model the enhancement was 6 percentage points. The results on average equivalence class size show that consonantal voicing resulted in substantial reduction in class size.

4. GENERAL DISCUSSION

Experiment 1 showed that consonantal voicing fails to disambiguate a majority of visually ambiguous words. At the

same time, the L+V transcriptions do substantially reduce the number of words in equivalence classes.

Experiment 2 showed that for the small set of keywords in the BKB sentence lists, L+V results in an increase in unique words over the LA transcriptions. However, the results of Experiment 2 did not account fully for the McGrath and Summerfield data, which showed greater enhancements with the acoustic F0 supplement in behavioral tests.

Several different factors could contribute to the McGrath-Summerfield results and [1,8], including effects due to syntactic and semantic levels of processing. However, we are intrigued with another possible contributor to word identification with F0, which is suggested by the foregoing analyses of word equivalence class size. Contemporary models of word recognition incorporate frequency weighted decision rules [19]. Under conditions of ambiguity, the decision rule results in the selection of the most frequent word. Thus, accurate word recognition can occur as a function of both perceptual uniqueness and lexical biasing. Under the lipreading with F0 condition, lexical biasing could resolve the remaining ambiguity of words, particularly in sentence sets designed to sample frequent words, as in the BKB sentences.

5. REFERENCES

1. Rosen, S. M., Fourcin, A. J., and Moore, B. C. J. (**1981**). "Voice pitch as an aid to lipreading," Nature **291**, 150-152.

2. McGrath, M., and Summerfield, Q. (**1985**). "Intermodal timing relations and audio-visual speech recognition by normal-hearing adults," J. Acoust. Soc. Am. **77**, 678-685.

3. Auer, E. T. Jr., and Bernstein, L. E. (**1996**). "Homopheneity in speechreading: Effects of phonemic equivalence classes on the structure of the lexicon," in *NATO-ASI Series F: Computer and Systems Sciences. Speechreading by Humans and Machines* edited by D. Stork and M. Hennecke (Springer-Verlag, Berlin).

4. Berger, K. W. (**1972**). "Visemes and homophenous words," Teacher of the Deaf **70**, 396-399.

5. Lisker, L., and Abramson, A. S. (**1967**). "The voicing dimension: Some experiments in comparative phonetics," in *Proc. Sixth International Congress of Phonetic Sciences* (Academia, Prague) 563-567.

6. Lehiste, I. (**1970**). *Suprasegmentals* (MIT Press, Cambridge, MA).

7. Cutler, A., and Butterfield, S. (**1992**). "Rhythmic cues to speech segmentation: Evidence from juncture misperception," J. Mem. Lang. **31**, 218-236.

8. Waldstein, R. S., and Boothroyd, A. (**1985**). "Speechreading supplemented by single-channel and multi-channel tactile displays of voice fundamental frequency," J. Speech Hear. Res. **38**, 690-705.

9. Bernstein. L. E. (**1995**). "Sequence comparison techniques can be used to study speech perception," in *Symposium on Speech Communication Metrics and Human Performance* (National Technical and Information Service: AL/CF-SR-1995-0023).

10. Raphael, L. J. (**1971**). "Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English," J. Acoust. Soc. Am. **51**, 1296-1303.

11. Hnath-Chisolm, T., and Kishon-Rabin, L. (**1988**). "Tactile presentation of voice fundamental frequency as an aid to the perception of speech pattern contrasts." Ear Hear. **9**, 329-334.

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

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

14. Seitz, P. F., Bernstein, L. E., and Auer, E. T., Jr. (**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).

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

16. Nusbaum, H. C., Pisoni, D. B., and 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).

17. Bernstein, L. E., Demorest, M. E., and Eberhardt, S. P. (**1994**). "A computational approach to analyzing sentential speech perception: Phoneme-to-phoneme stimulus-response alignment," J. Acoust. Soc. Am. **95**, 3617-3622.

18. Bench, J. and Bamford, J.(**1979**). *Speech Hearing Tests and the Spoken Language of Hearing-Impaired Children* (Academic, London).

19. Luce, P. A., Pisoni, D. B., and Goldinger, S. D. (**1990**)."Similarity neighborhoods of spoken words," in *Cognitive Models of Speech Processing*, edited by G.T.M. Altmann (MIT Press, Cambridge), pp 122-147.