EFFECT OF LEXICAL FREQUENCY AND NEIGHBORHOOD DENSITY ON AUDIOVISUAL SPOKEN WORD RECOGNITION

Kuniko Nielsen

Oakland University, USA
nielsen@oakland.edu

ABSTRACT

The current study investigated how lexical factors influence the intelligibility of spoken words, and how those effects interact with visual information. A forced-choice word-identification experiment was carried out under auditory-visual and auditory-only conditions with varying S/N ratios, and the effects of lexical frequency and neighborhood density on identification accuracy were analyzed. The results showed a significant interaction between frequency and modality: in auditory-only condition, words with high frequency (within the forced-choice options) were recognized more accurately, while in auditory-visual condition, words with low frequency were recognized more accurately. The effect of neighborhood density was also significant: words from dense neighborhoods were recognized more accurately than words from sparse neighborhoods.

Keywords: lexical frequency, neighborhood density, audiovisual speech perception

1. INTRODUCTION

It has been well established that aspects of lexical structure influence the way we produce and perceive speech. Words with higher frequency are more likely to be reduced than words with lower frequency, and words with lower frequency are more likely to be hyperarticulated than words with higher frequency [2]. At the same time, high frequency words are recognized faster and more accurately than low frequency words [3]. That is, attested hyperarticulation on low frequency words does not appear to facilitate speech perception. Lexical neighborhood density has also been shown to have similar effects on speech production and perception: words from dense neighborhoods are more likely to be hyperarticulated than words from sparse neighborhoods. For example, Goldinger and Summers [4] showed that the VOT difference in the initial stops of voiced-voiceless minimal pairs is greater for pairs of words from dense neighborhoods than for those from sparse neighborhoods. Wright [13] as well as Munson and Solomon [7] showed that words from sparse neighborhoods are produced with consistently greater vowel reduction than words from dense neighborhoods. The facilitatory effect of hyperarticulation on speech perception has been well attested (e.g., Bradlow, et al. [1]), and thus it would be predicted that hyperarticulation associated with words from dense neighborhoods would facilitate their perception. However, similar to the effect of lexical frequency, previous studies suggests the interaction to be more complicated. In Luce and Pisoni [6], the word recognition accuracy was higher among words with lower neighborhood density (=lexically easy) in the perceptual identification task. However, in their lexical decision task, the effect of neighborhood density was prominent only among low-frequency words, where accuracy was higher for words in high-density neighborhoods (=lexically difficult). Further, Vitevitch [11] reported fewer errors for words with many similar sounding neighbors than for words with few neighbors. In Vitevitch [12] speakers name equally familiar pictures more quickly and accurately when the picture name is from a dense neighborhood than when it is from a sparse neighborhood.

Note that the results of these word recognition tasks are inevitably affected by lexical effects from two sides: speaker and listener. Even if speakers hyperarticulate lexically difficult words (which would benefit listeners), the greater lexical competitions in the listener's lexicon could attenuate the potential facilitation caused by the hyperarticulation. If we could mitigate the competition in listeners' lexicon, it would enable us to test whether hyperarticulation associated with lexically difficult words would facilitate their perception more directly. For this purpose, a forced-choice word identification task was designed in which subjects were given three options in the form of *minimal triplets*. This paradigm enables us to minimize the (inhibitory) effect of lexical completion, because the

competition is limited to three words, whose neighborhood density counts are very close to each other.

In addition, we also aim to investigate how those lexical effects interact with the presence of visual information. It is well known that the presence of visual cues greatly increases the intelligibility of a speech signal (Sumby & Pollack [9]). If speakers hyperarticulate lexically difficult words, it would not only increase the acoustic cues but also the visual cues. Given the attested benefit of visual information in speech intelligibility, it would be predicted that the presence of visual cues would facilitate the perception of lexically difficult words more than that of lexically easy words.

2. METHOD

2.1. Subjects and stimuli

Sixteen native speakers of American English (11 females and 5 males, age 18-25) with normal hearing and normal or corrected vision served as subjects for this experiment. The material consisted of 108 English words that met the following criteria: (1) monosyllabic (CVC), (2) the initial consonant was one of the 15 American English consonants /p, t, k, f, θ, s, b, d, g, v, ð, ʃ, ʧ, r, w/, and 3) had a familiarity rating of 6.0-7.0 on the 7-point Hoosier Mental Lexicon scale (Nusbaum, Pisoni, & Davis [8]). The lexical frequency and phonological neighborhood density (B) were obtained from the Washington University in St. Louis Speech and Hearing Lab Neighborhood Database. The consonants were arranged into six triplet groups such that each set contains auditorily highly confusable consonants in noise: [p/t/k, b/d/g, f/θ/s, v/ðb, r/w/v, s/ʃ/ʧ]. Note that /b/, /s/ and /v/ were included in two triplets for the sake of forming suitable triplets. The six triplet groups were then classified into two groups, Easy [p/t/k, b/d/g, f/θ/s, v/ðb] and Hard [r/w/v, s/ʃ/ʧ]. The segments in the Easy group are expected to be easy to distinguish visually, while the segments in the Hard group are expected to be difficult to distinguish visually. The degree of visual confusability was determined based on [5]. Six minimal triplets (e.g., *pick, tic, kick*) were then chosen for each triplet group, yielding thirty-six minimal triplets (=108 words) in total.

The lexical frequency in the stimulus set was classified in the following manner. Within each minimal triplet, the word with the highest lexical frequency was labeled as "HI", and the word with the lowest frequency was then labeled as "LOW", and only HIs and LOWs were used as perception stimuli (=72 words). For example, for the first triplet in Table 1 [pick, tic, kick], only pick (=HI) and tic (=LOW) were presented in the experiment. The number of HIs and LOWs were balanced for each consonant. The same number of HIs and LOWs (36 HIs & 36 LOWs) were used as actual auditory or auditory-visual stimuli in the experiment, while all 108 words were presented on the screen (in the form of triplets) as the forced-choice response options. By comparing the intelligibilities between the groups HI and LOW, we hoped to determine the effect of lexical frequency on the rate of correct response.

Table 1: Example of perception stimuli.

HIs (the words with the highest frequency counts in the triplet) are shown in bold, and LOWs are shown in Italic. Numbers in parenthesis are lexical frequency and neighborhood density (B).

	Option 1	Option 2	Option 3
p/t/k	**pick** (55/34)	*tic* (3/28)	kick (16/30)
	pill (15/36)	till (50/35)	**kill** (63/36)
b/d/g	bowl (23/33)	*dole* (1/33)	**gole** (60/28)

The experimental stimuli were recorded in a sound booth in the UCLA Phonetics Laboratory. The speaker producing the stimuli was a trained phonetician. All utterances were recorded onto videotape, and then transferred onto a computer. The movie clips were edited into 72 three-second clips using Apple iMovie. The audio tracks were extracted from the edited movie files (sampling rate: 22000 Hz) so that Audio-Visual and Auditory-Only tokens had exactly the same sound tracks. Signal level was determined in terms of the peak RMS amplitude over a 30 ms window. All the audio files were then equated for the peak RMS amplitude at 80 dB (nominal). Noise was added to the speech at five levels of S/N: -10, -5, 0, 5, 10, 15 dB. S/N was defined by keeping the signal level constant at 80dB and manipulating the noise level from 65dB to 90dB. The noise used in this experiment is flat shape, band-pass filtered at 200-6500 Hz using Kay Elemetrics' MultiSpeech. The experimental audio stimuli were calibrated at a fixed 85 dB SPL for all the S/N ratios and presented binaurally over headphones. The visual displays were presented on a 20-inch computer screen.

2.2. Procedure

The stimuli were presented using Psyscope 1.2.5. Each subject was seated in front of a computer in a sound booth. A three-button button box was placed directly below the computer screen. Each main session was divided into two blocks: auditory-only (A) and auditory-visual (AV). In the A block, subjects listened to a word while they saw three response options (a minimal triplet) on the screen, and they were asked to press the button which corresponded to the word they thought they heard (forced-choice). In the AV block, the same subjects were asked to watch and listen to video clips of the speaker on the screen, again with a forced-choice response from three words, and to press the button which corresponds to the word they thought they heard. The subjects were told to guess when unsure, or if they could not detect the stimulus in the noise. Due to experimental time limitations, the 72 words were divided into two lists of 36 words each. The program randomized 36 words and 6 S/N ratio settings within a S/N setting and a block, respectively, and recorded both the key response and the reaction time. Each subject went through 432 total trials: 2 blocks (A and AV), 6 S/N ratio settings in each block, 36 trials in each setting. Each word was presented once for each trial, and one session lasted about 45 minutes, including the initial practice session. The order of blocks and the word lists within blocks were counterbalanced across subjects.

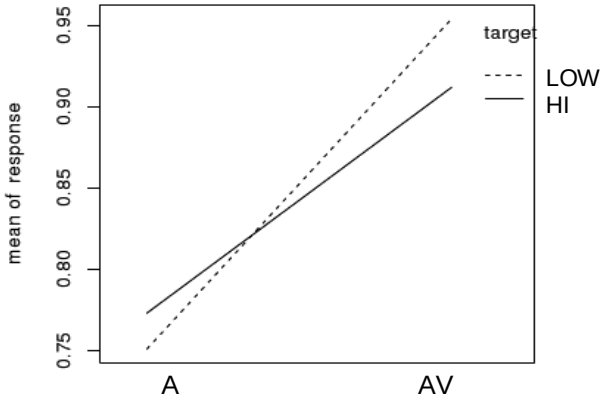
3. RESULTS

The independent variables in this study are 1) Signal-to-Noise ratio (S/N), 2) presence of visual information (A vs. AV), 3) categorical lexical frequency (HI vs. LOW), 4) log lexical frequency, 5) neighborhood density, and 6) visual confusability of the consonants (Easy vs. Hard). The effect of each of these variables on the correct response rate was analyzed by generalized mixed-effects modeling (using the glmer function in R).

As expected, S/N had a significant effect on the correct response rate [$z=12.448$, $p<0.001$], and so did the presence of visual information [$z=2.796$, $p<0.01$], where higher S/N as well as the presence of visual information (AV condition) increased the correct response rate. Although neither categorical lexical frequency (Hi vs. LOW) or log frequency had a significant main effect [$p>0.1$], the interaction between the presence of visual information and categorical lexical frequency was significant [$z=-2.152$, $p<0.05$]. That is, in the auditory-only condition, lexical frequency of the stimuli and the correct response rate had a positive correlation. On the other hand, in the audio-visual condition, the relationship between the two reversed and the words with lower frequency had higher correct response rate (Figure 1). The effect of neighborhood density was also significant [$z=2.341$, $p<0.05$]: words in dense neighborhoods showed higher correct response rate. As expected, the effect of visual confusability of consonants (Easy vs. Hard) was also significant [$z=-4.964$, $p<0.0001$. The presentation order and the word list did not have an effect [$p>0.1$].

Figure 1: Interaction between lexical frequency and presence of visual information.

The presence of visual information facilitated the perception of all stimulus items, but the effect was greater for LOW words.



4. DISCUSSION AND CONCLUSION

The results revealed that two lexical factors influence audiovisual spoken word recognition: lexical frequency and neighborhood density. A significant interaction was found between lexical frequency (classified into HI and LOW based on their relative frequency within the triplet) and modality, showing that the benefit of additional visual cues varied across the two frequency groups. Although the presence of visual cues greatly facilitated the identification accuracy for both groups, the effect was greater for low frequency words, which is lexically more "difficult" and thus expected to be (relatively) hyperarticulated by the speaker. In the auditory-only condition, words with high frequency were recognized more accurately, while in the auditory-visual condition, this pattern was reversed and words with low frequency were recognized more accurately. This interaction is predictable if we

consider the two independent forces at work: lexical competition and effect of visual cues. For a given minimal triplet, the degree of (auditory-based) lexical competition, an inhibitory force, should be the same for the auditory-only and auditory-visual conditions. On the other hand, the effect of visual cues is additive to the auditory cues, and thus the degree of facilitatory force is greater for the auditory-visual condition. Further, words with low frequency were predicted to have additional cues (both auditory and visual) as a consequence of hyperarticulation. In the absence of visual information, the inhibitory effect of lexical competition might have overpowered the facilitatory effect of hyperarticulation, resulting in higher accuracy for high frequency words, while in the presence of visual information, the facilitatory effect of hyperarticulation for low frequency words appears to have overpowered lexical competition, resulting in higher accuracy for low frequency words.

The results also revealed a significant effect of neighborhood density, where words in dense neighborhoods were recognized more accurately than words in sparse neighborhoods. This positive correlation between density and response accuracy suggests that the hyperarticulation associated with words from dense neighborhoods facilitates speech perception. Further, it also indicates that the effect of neighborhood density in perception is not necessarily inhibitory, which is in agreement with previous studies. Note that the range of neighborhood density was limited in the current study (i.e., all words were monosyllabic, and thus their density counts were relatively high): an additional study with more stimuli and a wider range of density is required to further elucidate the neighborhood density effect.

The results reported in the current study are preliminary, and call for further investigation in various directions. First, the current study involved only one speaker's production. Given the attested cross-speaker variability of speech production, it is important to replicate the same pattern of results with a larger number of speakers. Second, it is crucial to examine the correlation between quantitative measures of hyperarticulation (both acoustic and visual) and their perceptual accuracy in order to estimate the magnitude of perceptual facilitation. Lastly, a study by Tye-Murray et al. [10] suggests that visual neighborhoods exist, and that they affect auditory–visual speech perception. By adding visual neighborhood density as an additional factor, our data might reveal further information regarding the interaction of lexical effects and audio-visual speech perception.

5. REFERENCES

[1] Bradlow, A.R., Pisoni, D.B. 1998. Recognition of spoken words by native and non-native listeners: talker-, listener- and item-related factors. *Research on Speech Perception Progress Report* No. 22, 73-94.

[2] Bybee, J. 2001. *Phonology and Language Use.* Cambridge, UK: Cambridge University Press.

[3] Goldinger, S.A., Pisoni, D.B., Luce, P.A. 1996. Speech perception and spoken word recognition: research and theory. In Lass, N.J. (ed.), *Principles of Experimental Phonetics*, 277-327.

[4] Goldinger, S.D., Summers, W.V. 1989. Lexical neighborhoods in speech production: A first report. *Research on Speech Perception Progress Report* No. 15, 331-342.

[5] Jiang J. 2003. *Relating Optical Speech to Speech Acoustic and Visual Speech Perception*, Ph.D. Diss., UCLA.

[6] Luce, P.A., Pisoni, D.B. 1998. Recognizing spoken words: the neighborhood activation model. *Ear & Hearing* 19, 1-36.

[7] Munson, B., Solomon, N.P. 2004. The effects of phonological neighborhood density on vowel articulation. *JSLHR* 47, 1048-1058.

[8] 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 Speech Perception Progress Report* No. 10.

[9] Sumby, W., Pollack, I. 1954. Visual contribution to speech intelligibility in noise, *J. Acoust. Soc. Amer*. 26, 212-215.

[10] Tye-Murray, N., Sommers, M., Spehar, B. 2007. Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends in Amplification* 11(4), 233-241.

[11] Vitevitch, M.S. 1997. The neighborhood characteristics of malapropisms. *Language and Speech* 40, 211-228.

[12] Vitevitch, M.S. 2002. The influence of phonological similarity neighborhoods on speech production. *JEP: Learning, Memory, and Cognition* 28, 735-747.

[13] Wright, R.A. 2004. Factors of lexical competition in vowel articulation. In Local, J.J., Ogden, R., Temple, R. (eds.), *Papers in Laboratory Phonology VI*. Cambridge: Cambridge University Press.