Psychoacoustic measures of stop production in Cantonese, Greek, English, Japanese, and Korean

Tim Arbisi-Kelm*, Mary E. Beckman**, Eunjong Kong**, and Jan Edwards*
*University of Wisconsin-Madison, Madison, WI; **Ohio State University, Columbus, OH

INTRODUCTION AND RATIONALE

- Spectral analyses of stop bursts have revealed that place of articulation can be differentiated, using both invariant and time-varying cues present within the acoustic signal (e.g., Stevens and Blumstein, 1978; Forrest et al., 1988).
  - ➢ While prior studies have had some success in uncovering such cues for American English, it is not clear whether these parameters are equally pertinent to stop consonant classification in other languages.
  - ➢ Measures of burst frequency must be appropriate for both the compact and potentially bimodal peaks of velar stops and the diffuse bursts of alveolar stops, as well as capture the variability found across vowel contexts.
- Furthermore, one of the main limitations of linear acoustic analysis is that it imposes different scales of loudness and frequency on the acoustic signal than does the human ear, thus generating power spectra with different frequency distributions than are produced by the auditory system (e.g., Zwicker 1961; Kewley-Port, 1983).
- This study evaluates whether a spectral modes analysis would be more tractable if the spectrum is first "smoothed" by applying a psychoacoustic transform, so that peaks will correspond to audible concentrations of energy in the representation of band-specific loudness at the auditory periphery.

METHOD

Materials

- Languages: Cantonese, English, Greek, Japanese, Korean
  - All data recorded in each country with a native speaker as the experimenter.
- Participants:
  - ➢ 10 adult speakers of each language.
- Stimuli:
  - ➢ Velar and alveolar stop consonants placed in word-initial position in familiar words in the following vowel contexts: /ɑ, e, i, o, u/.
  - ➢ Three word forms for each vowel context.
- Word repetition task: Participant asked to repeat word, given auditory prompt.

Acoustical analysis

Analysis 1: Spectral moments analysis

- Using Praat (Boersma & Weenink, 2001), we first downsampled the audio files from 44 to 20 kHz in order to mimic the sampling rate used in Forrest et al. (1988), and high-pass filtered the sound files at 70 Hz to minimize outside noise.
- We then generated 512-point fast Fourier transforms (FFT) for each token across a 20-ms Hamming window, centered at the burst, in order to obtain a frequency distribution of the burst energy.
- The resulting long-term averaged spectra were then normalized and converted into probability distributions, in order to compute the linear frequency scale spectral moments.
- The first four spectral moments—centroid, standard deviation, skewness, and kurtosis—were then calculated, using the formulas defined in Forrest et al. (1988).

Analysis 2: Auditory-based analysis

- Spectral slices were generated across a 10-ms Hamming window, centered at the burst, to obtain a frequency distribution of the burst energy.
  - ➢ The very small window was used to effectively isolate the front cavity resonances of the burst, and thus minimize influence of the following vowel.
- We first designed a method of calculating specific loudness (SL) against equal rectangular bandwidth (ERB) as an initial step in developing psychoacoustic measures of loudness.
- We created a function to transform long-term averaged spectra (Ltas) from dB into sones (specific loudness), using programs modeled after those used in Moore, Glasberg, and Baer (1997).
- Three measures were developed and calculated for each burst spectrum:
  - ➢ The highest amplitude frequency ("peak ERB"): used to estimate the length of the front cavity, and thus the point of constriction during production of the target consonant.
  - ➢ The proportion of energy within the most prominent spectral peak ("compactness index"): analogous to the second spectral moment of linear analysis (i.e., standard deviation), in order to parameterize the spectral energy distribution along the compact/diffuse dimension.
  - ➢ A second compactness measure ("range ERB"): the ERB bandwidth contained within a 3-sone drop from the peak amplitude frequency.

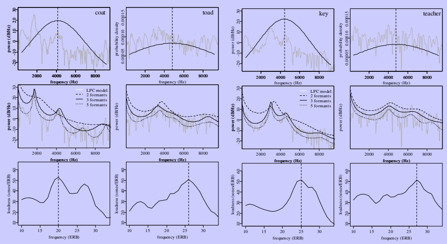

Figure 1. Representations and models of burst spectra in English /k/ and /t/ before a back vowel in the words *coat* and *toad* (left) and before a front vowel in the words *key* and *teacher* (right). Top row= FFT spectra (in gray) with a Gaussian distribution centered at the first spectral moment overlaid in black. Middle row= LPC models specified to identify two, three, or five "formant" peaks. Bottom row= psychoacoustic spectrum (sones against ERB) with cursors at the loudest peak.

Acoustic measure	Description and prediction
1) peak amplitude frequency (peak ERB)	- acute/grave dimension - the peak amplitude frequency, representing the point of highest specific loudness (measured in sones) - higher frequency peaks= shorter front cavity (alveolars, front velars) - lower frequency peaks= longer front cavity (back velars)
2) compactness index (CI)	- compact/diffuse dimension - from a normalized spectrum, the proportion of energy (in sones) within a 3-ERB band centered at peak amplitude frequency - higher value= compact frequency peak (velars) - lower value= diffuse frequency peak (alveolars)
3) ERB mode bandwidth (range ERB)	- compact/diffuse dimension - ERB range (high-low) contained within a 3-sone drop from the peak amplitude frequency - lower value= compact frequency peak (velars) - higher value= diffuse frequency peak (alveolars)

Table 1. Acoustic measures developed for the auditory-based analysis.

RESULTS

Linear discriminant analyses

- Spectral moments analysis: for the English stops, it was found that the three measures correctly predicted the place of articulation category 71% of the time. The measures were less successful in discriminating place of articulation for stops in the other four languages.
- Auditory-based analysis: the measures peak ERB, CI, and range ERB distinguished alveolar and velar stops for all five languages with a higher degree of accuracy overall than did the linear spectral moments measures.

	English	Cantonese	Greek	Japanese	Korean
Spectral moments	.71	.66	.62	.59	.62
peak ERB + range ERB	.81	.69	.69	.73	.76
peak ERB + CI	.86	.74	.76	.84	.79

Table 2. Overall proportions of correctly-predicted stops for each linear discriminant analysis, compared across languages.

.

Differences across languages and vowel contexts

- For English, peak ERB + CI outperformed spectral moments measures in correctly categorizing alveolars in each vowel context (including alveolars overall, .96 vs. .76), as well as overall for velars (.73 vs. .65), but was not more successful in classifying front velars (/e/= .55 vs. .52; /i/= .53 vs. 63).
- This was not the case for Cantonese, Greek, and Japanese, where peak ERB + CI performed equal to or better than spectral moments measures in all vowel contexts (including front vowels), with the exception of Cantonese /a/. (Both auditory-based and spectral measures both poorly discriminated Korean front velars.)
- While velars on the whole were predicted with greater accuracy by the auditory-based measures, the alveolar results were even stronger. Peak ERB + CI classified alveolar stops with considerably greater success than did spectral moments in Korean (.91 vs. .64), Japanese (.78 vs. .37), Greek (.83 vs. .61), and Cantonese (.77 vs. .65).

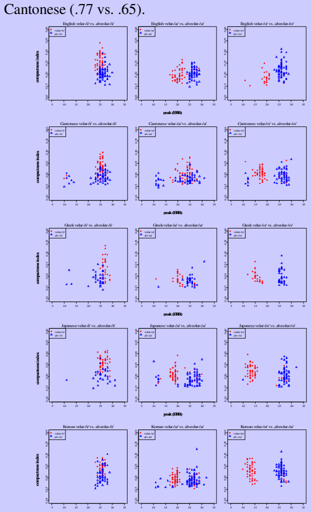

Figure 2. Scatterplots of alveolar and velar stop bursts before three vowels /a,i,o/ in all languages, combined for all subjects (Korean stimuli included /o/ in place of /u/ contexts). Compactness index values are plotted along the y-axis, and peak amplitude frequency values are plotted along the x-axis.

- The scatter plots in Figure 2 show that the peak amplitude frequency is the relevant parameter for distinguishing stop category before the non-front vowels, while the compactness index separates velar and alveolar tokens in the front vowel contexts, where peak ERB values are similar across all tokens.
- Although the results of the linear discriminant analysis revealed that velars were less consistently identified in front vowel velar contexts, the scatter plots in Figure 2 nevertheless show a fairly clear separation between alveolars and velars before front vowels.

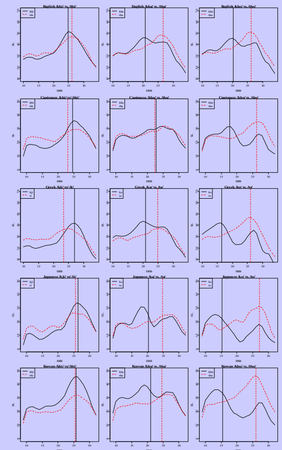

Figure 3. Average SL/ERB spectra of all adult subjects for each language and across three vowel contexts, comparing alveolar (red) with velar (black) stops. Vertical lines indicate peak amplitude frequency.

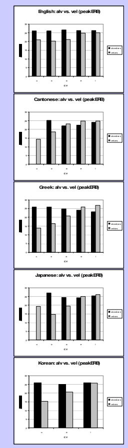

Figure 4. Comparison of mean peak amplitude frequency of velar bursts before each vowel context, collapsed for all subjects and separated by language.

- In all languages, peak ERB values were higher in alveolars than velars in back vowel contexts.
- Back velars had considerably lower peak ERB values in Greek (/u/= 13.8, /o/= 16.4), Japanese (/o/= 19.3), Cantonese (/u/= 14.3, /o/= 18.7) and Korean (/u/=15.2) than in English (/u/= 26.2, /o/= 26.2).
- In all languages except English, there was a gradient effect of vowel context on velar peak ERB values, which gradually increased as the following vowel became more front.
- In English, the influence of following vowel context on velar peak ERB values was more categorical than gradient, with one value for back vowels and another value for front vowels.

DISCUSSION AND CONCLUSION

- Traditional spectral analyses based on linear model scales successfully distinguished back velar stops from alveolar stops in English, but were far less successful in discriminating velars from alveolars in front vowel contexts in English, and across all vowel contexts in Cantonese, Greek, Japanese, and Korean.
- Auditory-based measures had greater success in distinguishing velar and alveolar stops across all vowel contexts in all languages.
  - ➢ In back vowel contexts, alveolars and velars differed primarily with respect to peak amplitude frequency.
  - ➢ In front vowel contexts, alveolar and velar stops differed primarily in terms of the compactness of their amplitude frequency peaks.
- Thus, it appears that the predictive success of these two auditory-based measures is in their complementarity:
  - ➢ peak ERB discriminates velars from alveolars where they are most dissimilar along the acute/grave dimension — i.e., before non-front vowels.
  - ➢ the compactness index is most useful in distinguishing the two stop types in front vowel contexts, where the alveolars and velars still differ in terms of tongue posture.
- For all languages relative to English, velar stops were produced in more extreme back positions before back vowels.
  - ➢ Moreover, in all languages except English, there was evidence of a gradient realization of tongue backness in velars across vowel contexts, as indicated by the systematic variation in the ERB frequency of the loudest peak.

ACKNOWLEDGEMENTS

Supported by NIDCD grant R01 DC02932 to Jan Edwards

- Thanks to Hyunju Chung, Junko Davis, Fangfang Li, Sarah Schellinger, Laura Slocum, Asimina Syrika, and Junko Davis for their work on data collection, native-speaker transcription, and event-marking.
- Thanks also to the children who participated in the study, the parents who gave their consent, and the schools who let us use their facilities for testing.

the παιδολογος project
cross-language investigation of phonological development

156th Meeting of the Acoustical Society of America, Miami, Florida, November 10 – November 14, 2008