PERCEPTION OF ORAL RELEASE RATE FOR INITIAL VOICED STOPS

David R. Williams
Sensimetrics Corporation, Cambridge, MA 02139, USA

ABSTRACT

Listener preferences for variations in oral release rate (ORR) of initial labial, alveolar and velar voiced stops were tested. Three five-member continua were synthesized using a set of articulatory parameters to control the increase in oral cross-sectional area at stop release. For each continuum, naive and experienced subjects judged the typicality of each member against all others. Faster ORRs were preferred for labial stops than for alveolar and velar stops.

INTRODUCTION

Because of differences in mass of the primary articulator and contact length, initial rates of increase in cross-sectional area of the oral constriction may differ for stops made at the various places of articulation. Based on theoretical considerations, acoustical analyses and some physiological data [1, 2], faster rates of increase in cross-sectional area are expected for stops formed at the lips than for those formed by the tongue tip or by the tongue body. These differences in oral release rate (ORR) have acoustic and aerodynamic consequences which may play an important role in the perception of stop place of articulation.

This paper examines the perception of syllable-initial voiced stops that were synthesized with differences in ORR. The synthesis was achieved by means of a new approach which employs a set of articulatory parameters to control a Klatt synthesizer [3]. The purpose of the experiments was twofold. First, there was an interest in assessing the approach itself to determine the precision with which control parameters must be specified in order to achieve acceptable synthesis of certain phonetic categories. Second, it was of interest to examine listeners' ability to discriminate within-category variations for a previously untested articulatory variable.

THE HL SYNTHESIS APPROACH

The underlying motivation for the synthesis approach described here is to combine the simplicity of control that characterizes articulatory approaches to synthesis with the accuracy and computational efficiency of traditional formant synthesis [4, 5]. This hybrid approach employs a small set of high-level (HL) parameters to construct an acoustic-articulatory utterance specification which is then transformed by means of mapping relations into a specification in terms of the larger set of lower-level (LL) acoustic parameters needed to control a KLSYN88 formant synthesizer [6]. In effect, the HL synthesis system is a formant synthesizer preprocessor.

The HL synthesis parameters and their functions can be described in terms of three broad classes:

Class 1 parameters control the first four natural frequencies of the vocal tract (***f1***, ***f2***, ***f3***, ***f4***) and the fundamental frequency (***f0***). These parameters specify the configuration and slow movements of articulators which determine the global shape of the vocal tract.

Class 2 parameters control the cross-sectional areas of local constrictions formed by the lips (***al***) and by the tongue tip or blade (***ab***). They specify the fast movements of primary articulators that rapidly decrease/increase airflow within the oral tract.

Class 3 parameters control the cross-sectional areas of the glottal orifice (***ag***) and velopharyngeal port (***an***) and the pharyngeal volume (***ue***). They are used to specify opening/closing movements of the glottis and velum as well as active expansion or contraction of the pharynx.

The first step in determining values for the LL parameters is to calculate the pressures and flows at the supraglottal and glottal orifices using an aerodynamic model [7]. In addition to the Class 2 and 3 parameter values, inputs to the model include *agx*, the glottal orifice area as modified by supraglottal forces, and *acx*, the smallest supraglottal constriction area. The output of the model is the intraoral pressure (P_m) which, along with the orifice areas and a constant subglottal pressure (P_s) value, provides a basis for computing the LL source amplitudes **AV**, **AH** and **AF**.

Other settings and modifications of the LL parameters result from values of HL parameters specified by the user. In general, the Class 1 parameters are mapped directly to their corresponding LL parameters when the glottal area is modal and the velum is closed. An increased glottal area *agx* affects the LL formant bandwidths and the values of **OQ** and **TL**. The presence of voicing in the synthesis signal is conditional on *agx* and the calculated transglottal pressure drop (**AV** = 0 when $agx > 13$ mm^2 or $P_s - P_m < 3$ cm H_2O). Place-specific filtering of the frication is determined from a look-up table when **AF** > 0 based on the values of ***f2*** and ***f3***.

Of particular interest for the current study are rules that affect the value of **F1**. Specifically, the HL first natural resonance frequency is modified (*f1c*) when a class 2 parameter specifies a local (labial or alveolar) constriction to reflect the fact that the constriction is currently controlling its value. The value of *f1c* is approximated as the lowest frequency of a Helmholtz resonator with constriction area *acx* and with a constriction length and pre-constriction volume that are determined by the place of articulation. When the natural frequencies specify a tongue dorsum (e.g., velar) constriction, *acx* directly reflects the value of ***f1***.

[Other rules that are operative when the velopharyngeal port is open or there is lateralization or retroflexion of the tongue are not discussed here.]

EXPERIMENT

The purposes of the experiment were to (1) determine the range of acceptable ORR values, and (2) to examine the range of listeners' preferences for various ORRs in voiced stops with different places of articulation. Although ORR is acoustically a very complex variable, it is simulated here by changes in a single HL parameter which controls the rate of increase in area of an oral constriction.

SYNTHESIS

HL Input Parameters

The synthetic stimuli were modelled on three /ba/, /da/, and /ga/ utterances produced by a male talker. All stimuli were 300 ms in duration and had an ***f0*** contour which fell from 115 Hz after the burst to 85 Hz at the end of the vowel. During the vowel's steady-state portion, the natural frequency values were 700, 1150, 2400, and 3500 Hz. Initial ***f1***, ***f2*** and ***f3*** transitions were derived from the spoken utterances, and thus, were place-specific. The glottal area ***ag*** was constant at a (modal) value of 4 mm^2 throughout the stimuli.

Three five-member series of CV syllables were synthesized by varying the rate of increase in the area of oral constriction at stop release. For the labial and alveolar series, this entailed setting the slope of, respectively, the **al** and **ab** parameter trajectories to correspond to ORRs of 10, 15, 25, 50 and 100 cm^2/s. For the velar stops, the value of ***f1*** was manipulated so as to achieve *acx* values that corresponded to ORRs of 10, 15, 20, 30 and 50 cm^2/s.

Output Parameters

The effect of decreasing ORR was to increase the number of frication (**AF**) frames in the burst and to decrease the initial aspiration (**AH**) level. For the slower ORRs, voicing (**AV**) onset was not simultaneous with oral release, but was delayed by as much as 15 ms due to high intraoral pressure. Initial changes in **OQ** and **TL** were much more abrupt for the faster ORRs than for the slower ones.

METHODS

For each place of articulation series, stimulus trials were constructed by pairing each ORR with every other ORR, including itself. Five randomized blocks of the 25 distinct stimulus pairs in each series were recorded on audio tape for presentation. The stimuli were presented over headphones in a quiet room.

Subjects were asked to compare the two stops in a pair and indicate whether the first or second member was the better exemplar of the particular voiced stop. An answer was requested on each trial, even if it was thought only to be a guess.

SUBJECTS

A group of fourteen "naive" subjects were recruited from the local academic community. A short questionnaire filled out prior to testing revealed that all subjects spoke only English, and none had a history of hearing impairment. The subjects were paid for their participation.

Twelve "experienced" subjects, all volunteers from the MIT Speech Comm. Group, were also tested. This subject group included senior graduate students, postdoctoral researchers and the author. All were native speakers of English who were familiar with phonetics and had experience judging synthetic speech.

RESULTS

Although all members of a series sounded like exemplars of that stop type, there were clear differences among the stimuli within a series. Most notably, at the slowest ORR, all stops sounded somewhat voiceless and even fricated. In the alveolar and velar series, stimuli with the fastest ORRs had a somewhat intrusive /y/-glide following the stop.

Table 1. Normalized scale values for 14 "naive" subjects at each ORR. Second ORRs are for the velar series.

Normal Scale Values	ORR in cm^2/s				
	10	15	25 /20	50 /30	100 /50
/ba/	-0.99	-0.38	0.12	0.54	0.71
/da/	-1.12	-0.12	0.73	0.30	0.22
/ga/	-0.44	0.18	0.00	-0.18	0.44

Naive subjects

Table 1 shows normalized scale values for the naive subjects. The scores were computed by first transforming the percentage of times that the indicated stimulus was preferred over each other member of a series (in both positions) to normalized (z-score) units and then summing over scores (see [8]). The table shows the average normalized scale values for the pooled data. (A value of 0 represents no preference for or against.)

The scale values indicate that the naive subjects preferred the two fastest ORRs for the labial stops and the intermediate (25 cm^2/s) ORR for the alveolars. There was clearly a negative preference for the slowest ORR. For the velars, it would appear that the fastest ORR was again preferred. However, the range of scale values here are small, and the raw percentage scores did not suggest any strong preferences.

Experienced subjects

Table 2 shows the results for the 12 experienced subjects. Like the naive subjects, this group preferred the intermediate ORR for the alveolars and the faster ORRs for the labials. The lower labial scale values and spread in preference to include the intermediate ORR (/ba/) are due to the fact that three

subjects preferred the slower ORRs over the two fastest ORRs (cf. /ba/-9). Nevertheless, there was overall a strong negative preference for the slowest labial and alveolar ORRs.

Table 2. Normalized scale values for 12 "experienced" subjects at each ORR. Second ORRs are for the velar series.

Normal Scale Values	ORR in cm^2/s				
	10	15	25 /20	50 /30	100 /50
/ba/-9	-1.06	-0.30	0.31	0.43	0.62
/ba/	-1.11	-0.14	0.39	0.36	0.50
/da/	-0.51	-0.29	0.82	0.26	-0.86
/ga/	-1.05	-0.26	0.17	0.75	0.39

Unlike the data in Table 1, these data reveal a strong preference for the velar stops with the 30 cm^2/s ORR and a strong preference against the slowest velar ORR. It is also notable that the experienced subjects were much less tolerant of the /y/-glide following the release of the aveolar stop with the fastest ORR.

DISCUSSION

The results show that intermediate ORRs were preferred for all places of articulation, and that faster ORRs were also preferred for the labial stops; slower ORRs were generally not preferred. It is thus possible that a single ORR (ca. 30-40 cm^2/s) could be used to synthesize all voiced stops. Although a range of ORRs were found to be acceptable within each stimulus series, the gradient of scale values differed as a function of stop place of articulation.

CONCLUSIONS

The present findings demonstrate listeners' ability to discriminate within-category phonetic variations, and thus contribute to the ongoing discussion of phonetic prototypes. As the stimuli used here were all articulatorily plausible, the study represents a refinement on previous assessment techniques. Relatedly, these results and others [9, 10] suggest that the scope of potential prototype definitions might reasonably be expanded to embrace speech production as well as perception. [Research supported in part by a grant from NIH.]

REFERENCES

[1] Fant, G. *Speech Sounds and Features*. Cambridge: MIT Press, 126.

[2] Stevens, K. N. (forthcoming) *Acoustic Phonetics*.

[3] Stevens, K. N., and C. A. Bickley (1991) "Constraints among parameters simplify control of Klatt formant synthesizer." *J. Phonetics 19*, 161-174.

[4] Stevens, K. N., C. A. Bickley, and D. R. Williams (1994) "Control of a Klatt synthesizer by articulatory parameters." *Proceedings 3rd Int'l. Conf. Spoken Language Processes*, Yokohama, Japan.

[5] Williams, D. R., K. N. Stevens, and C. A. Bickley (1992) "Inventory of phonetic contrasts generated by high-level control of a formant synthesizer." *Proceedings 2nd Int'l. Conf. Spoken Language Processes*, Banff, Alberta, Canada, 571-574.

[6] Klatt, D. H, and L. C. Klatt (1990) "Analysis, synthesis, and perception of voice quality variations among female and male talkers." *JASA. 53*, 1070-1082.

[7] Stevens, K. N. (1993) "Models for the production and acoustic of stop consonants." *Spch. Comm. 13*, 367-375.

[8] Green, B. (1974) "Paired comparison scaling procedure." In G. M. Maranell (ed.), *Scaling: a sourcebook for behavioral scientists*. Chicago: Aldine. Pp. 93-97.

[9] Williams, D. R. (1994, June) "Modelling changes in magnitude and timing of glottal and oral movements for synthesis of obstruent consonants." *JASA 95*, 2815 (A).

[10] Williams, D. R. (1994, November) "Perception of fricatives synthesized by higher-level control of a Klatt synthesizer." *JASA 96*, 3227 (A).