

HIGHER-LEVEL SYNTHESIS OF DEVOICEABLE SYLLABLES

Jonathan Rodgers

Institute of Phonetics and Digital Speech Processing, University of Kiel, Germany

University of Cambridge Department of Linguistics, U.K.

ABSTRACT

Structural segmental properties offer a partial account of patterns of devoicing in English vowels. HLsyn is used to test the validity of explanations in terms of articulatory-acoustic relations that are thought to underlie realizations found in English and German. Manipulation and appropriate phasing of parameters corresponding to subglottal pressure, degree of tongue-blade constriction and degree of glottal opening generate a range of convincing realizations that mirror those found in genuinely natural speech.

1. VOWEL DEVOICING IN ENGLISH

[1, 2] established that rather than a simple by-product of universal biomechanical constraints, vowel devoicing is a complex Connected Speech Process. The devoicability of a vowel is determined by segmental properties that affect aerodynamics, but also by metrical and lexical factors and the demands of potentially competing speaker-specific strategies. The factors conditioning vowel devoicing are primarily aerodynamic, and putatively universal, and such aerodynamic considerations stemming from segmental properties of the syllable can account for realizations spanning a range of voicing possibilities: for example, the syllable of interest in *article* may be realized with a voiced tap [ɾ], a heavily fricated release [tʰɪ] or a voiceless vowel [t̥ɪ].

2. USING SYNTHESIS TO EXAMINE ARTICULATORY-ACOUSTIC RELATIONS

The study reported here seeks to explore (i) the importance of lowered subglottal pressure and lax articulatory setting believed to be concomitant with unstressed syllables; (ii) the trade-off between tongue-blade constriction and vocal fold adduction in generating realizations with varying degrees of frication and aspiration. HLsyn offers an accessible model for testing articulatory and aerodynamic hypotheses, and is more appropriate for investigating natural speech than an invasive physiological study: in pilot trials both fibre-optic nasendoscopy and electroglottography inhibited the natural casual speech that is of interest here.

HLsyn is a quasi-articulatory synthesizer whose design is based on the observation that the values of the 40-odd parameters used to control Klatt-type synthesizers are not independent, but subject to inter-parameter constraints. The small set of high-level parameters which captures these constraints is more closely related to the actual states and articulatory movements in the vocal tract than are lower-level Klatt parameters. HLsyn uses a set of mapping relations to transform the HL parameters into the values of the corresponding lower-level KLSyn parameters [3, 4].

3. EXPERIMENTS

A pilot study and two experiments are reported. The pilot study attempts to synthesize a word containing a devoiceable vowel by altering only fundamental frequency f_0 , formants f_1 – f_4 , tongue-blade constriction ab and glottal area ag . The experiments examine the effect of changes in subglottal pressure and compliance of the vocal tract walls and vocal folds; and the effect of varying glottal and oral constrictions.

The tokens produced were not subject to formal perceptual testing. During development there were some acoustic analyses, mainly with reference to spectrograms, but all evaluation was impressionistic, i.e. auditory only. At each stage of development, a set of tokens was played under good listening conditions to at least five colleagues in the Phonetics Laboratory in the University of Cambridge Department of Linguistics. All listeners were familiar with synthetic speech. They were free to listen to the set of tokens as many times as they wished, and in any order. They offered impressionistic comments only, and also selected what they considered the most natural token.

3.1. Synthesis of a word containing a devoiceable vowel

The pilot study examines how readily a devoiceable syllable occurs in a token created by careful copy-synthesis: it aims to see whether devoicing “falls out of” the HLsyn synthesis system.

A natural but careful realization of the word *article* spoken in a carrier sentence by a male subject was chosen as the item to be copied, as the /tɪk/ syllable in this word was frequently fully voiceless in the production experiments reported in [1, 2]. A spectrogram of the utterance is shown in Figure 1.

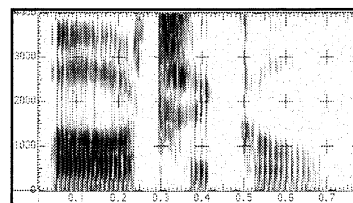


Figure 1. Spectrogram of naturally spoken *article*.

The /tɪ/ of the natural token has a frication phase of 55ms, an aspiration phase of 20ms, and a voiced phase of 45ms (4 cycles of periodicity); for segmentation criteria see [5]. The proportion of voicelessness, relative to the duration of the whole vowel, is 0.63, so the vowel is not especially devoiced by comparison with, for example faster and more casual realizations, where the vowel might be entirely voiceless. Waveforms, spectrograms and spectra in xwaves/ESPS were used to measure durations, f_0 and

frequencies of F1–F4 in steady state and transitions, and the values were input to Hlsyn.

Figure 2 shows a spectrogram of the synthesized token, which has a frication phase of 70ms, an aspiration phase of 10ms, and a voiced phase of 50ms (4 cycles of periodicity). The proportion of voicelessness in the critical syllable is 0.62.

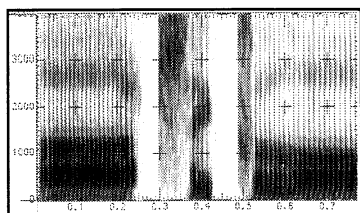


Figure 2. Version 1: spectrogram of synthesized token of *article*.

Speakers perceived the token in Figure 2 — henceforth Version 1 — as *article*, although it did not sound natural. In the vowel of /tik/ the proportion of voicelessness is the same as that in the natural token, but the durations of frication, aspiration, and voicing are all different. The copy was not intended to be exact, and was acceptable to all speakers, so no attempt was made to achieve the same durations as the natural token. However, it is clear from the spectrogram that the natural and synthesized tokens must sound very different: the periodicity throughout Version 1 is very regular, the /t/ burst sounds “hissy” and there is no decay of amplitude in the [i], which is indeed the least convincing part of the token.

Nonetheless, this example demonstrates the fundamental principle of Hlsyn, *viz* that the sequencing of silence, transient, frication, aspiration, and periodic excitation arise due to relative cross-sectional areas at the glottis and a point in the oral cavity, in this case the alveolar ridge. An acceptable if robotic token including a partially devoiced vowel can be generated with appropriate values for fundamental frequency, formants, and (appropriately phased) tongue-blade constrictions and vocal fold adduction-abduction.

3.2. Subglottal pressure and compliance of the vocal-tract walls and vocal folds

This experiment examines the influence on the devoiced syllable of two of three time-varying parameters added to Hlsyn in late 1997, after the first part of this research, reported in Section 3.1. The parameters control subglottal pressure and the compliances of the vocal-tract walls and vocal folds. In version 2.2 of Hlsyn the algorithm for calculating F0 is also altered: rather than being simply a function of the input f0 parameter values, the actual F0 is affected by subglottal pressure, vowel height, and tissue compliances.

ps (subglottal pressure) affects voiced segments, by altering F0 and AV (Klsyn amplitude of voicing), and is used to increase or decrease vowel amplitude for stressed and unstressed syllables. Version 1 (the token generated in Section 3.1) is divided into a stressed and an unstressed phase, the division being made at the point when closure for /t/ is achieved (i.e. $ab=0\text{mm}^2$).

The default value for ps is 8cm H₂O. [6, 7, 8, 9] were consulted to suggest a range of values for the stressed and unstressed sections of the word: for the stressed phase 8–12cm H₂O; for the unstressed phase 2–8cm H₂O. In Version 1, stressed syllables with ps greater than 10cm H₂O sound too loud, and unstressed syllables with ps less than 4cm H₂O sound faint. The best combination of values found was 9cm H₂O for the stressed, and 5cm H₂O for the unstressed phase. Furthermore, the difference between the values for ps in stressed and unstressed syllables must not be too great: in longer stretches of speech, however, a greater range of values may be appropriate.

dc, which represents change in the compliance of the vocal tract walls and vocal folds, can be used to simulate a more or less fortis/lenis utterance. [10] report that the compliances of the vocal tract walls change significantly throughout an utterance; the compliance of the vocal folds can also change: the main effect of such variations is to alter f_0 . The effect of a range of values from 0–50% in the unstressed phase of Version 1 was assessed. Values of greater than 40% sound “like cotton-wool”, whereas tokens with values below 20% are indistinguishable from those with no change; a value of 30% contributes to a more natural sounding token.

The ideal values of subglottal pressure and delta compliance suggested by independently altering ps and dc are applied to Version 1 to create Version 2, of which a spectrogram is shown in Figure 3. It has the same durations for frication, aspiration and voicing as the original *article* but sounds considerably more natural. Indeed Figure 3 not only sounds but also looks more like the natural token in Figure 2, especially with regard to the relative amplitude and spectral shaping of noise-excited segments, and the relative amplitude of all syllables

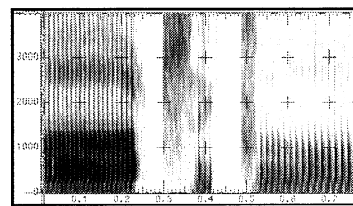


Figure 3. Version 2 (=Version 1 (Figure 2) using ps and dc).

3.3. Effect of oral and glottal constrictions

[1, 2] identified in English data a range of possible phasings of oral and glottal articulations that were also mirrored in data from German spontaneous speech. This experiment seeks to explore the interaction of oral and glottal constrictions and their effect on devoicing. It aims to investigate over what ranges of ab and ag frication will occur and over what ranges aspiration will occur.

This experiment has two parts: the first — Version 3 — seeks to synthesize a long phase of frication by prolonging tongue-blade constriction (ab); the second — Version 4 — seeks to synthesize a long phase of aspiration by delaying vocal fold adduction (ag). The starting file for both parts is Version 2: it has a /tik/ syllable with 70ms of frication, 10ms of aspiration, and 50ms of voicing (3 cycles of periodicity). The proportion of voicelessness in the critical syllable is 0.62.

The default value of **ab** is 100mm², and in the synthesis of Versions 1 and 2 **ab** was decreased from and increased to this value in the regions excluding the /t/ closure. For Version 3, the fricated token, an open value of 50mm² is used, simulating an articulation with only a small opening gesture.

A range of closing and opening rates was subjected to auditory analysis: closure was achieved over 10–40ms in 5ms intervals, and opening over 20–100ms in 10ms intervals. However, the most acceptable token came from using the same rate of closure as in Version 2 (and Version 1), i.e. 20ms for closure, 70ms for opening. The best percept of a fricated release came from increasing **ab** to a maximum of 25mm²: values of **ab** greater than 30mm² sound “hissy”, whereas values below 20mm² were “whistly”. The opening gesture sounded more natural when split into two phases: a steep rise to an intermediate value of 15mm² followed by a slower opening to 25mm² created the best burst.

The main effect on the /t/ of the change in **ab** is to replace the 10ms of aspiration in Version 2 with frication, creating a fricated phase of 80ms. There are still 50ms of voicing, as for Version 2. A further effect of the smaller opening (25mm² rather than 100mm²) of **ab** is to generate a shorter and better /k/ with slightly lower amplitude. The proportion of voicelessness is still 0.62.

The default value of **ag** is 4mm², which generates modal voicing given appropriate values of other parameters. [11] suggests a value of 30mm² for aspiration of stops and fricatives, appropriately phased with the formation and release of the constriction.

In Version 2, **ag** falls from 30mm² to 4mm² over 12ms, modal voicing coinciding with the open value of **ab**. A longer phase of aspiration can be generated by delaying the return of **ag** to modal value. A range of stimuli was generated which preserved the same rate of glottal abduction, but delayed the adduction movement relative to the release of the stop represented by the increase in **ab**: **ag** fell from 30 mm² to 4mm² over the same 12ms period ending 10, 20 or 30ms after **ab** reached 100mm². The most natural sounding token has voice onset delayed by 20ms relative to the original token: it has a frication phase of 35ms, aspiration phase of 60ms, and 25ms of voicing (only 1 period).

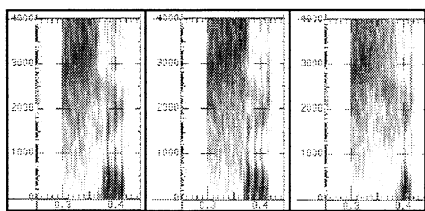


Figure 4. Spectrograms of (l. to r.) Version 2, Version 3 (long frication phase from altering **ab**) and Version 4 (long aspiration phase from altering **ag**).

Figures 4 and 5 show the /t/ syllable and oral pressure in Versions 2–4. Oral pressure remains high (3.5cm H₂O at 350ms) in Version 3 during of the fricated release of /t/, and prevents onset of voicing. In Version 2 and 4 oral pressure lowers more rapidly after release of the stop, but in Version 4 abducted vocal folds inhibit resumption of voicing.

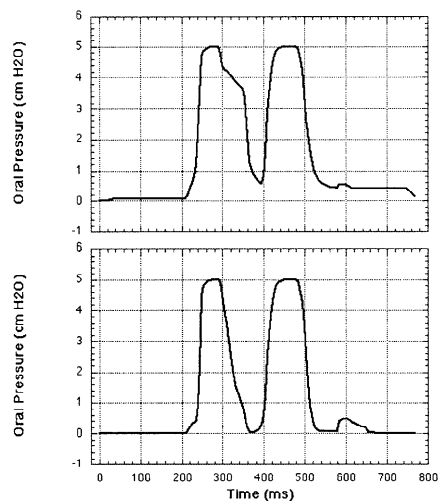


Figure 5. Oral pressure in tokens of *article* varying in **ab** and **ag**. Top is Version 3 (long frication phase, cf. Figure 4), bottom is Version 2 (cf. Figure 3) and Version 4 (long aspiration phase, cf. Figure 4).

4. CONCLUSIONS

HLsyn is a powerful research tool that allows useful insights into articulatory and aerodynamic processes; this study offers a limited introduction. Devoicing occurs in careful copy-synthesis, and the quality of the token can be improved by changes in parameters affecting aerodynamic and source properties. HLsyn’s articulatory parameters were used to explore putative strategies for realizing tokens found in English and German data, and the pressure-flow feature was used to examine oral pressure in stops where either oral constriction or glottal opening was maintained, either of which fosters devoicing.

Voicing in synthesis is a two-edged sword. On the one hand the naturalness and intelligibility of speech synthesis algorithms are enhanced by incorporating the findings of coarticulatory studies [12], particularly those concerned with voicing; on the other, intelligible synthetic speech can nonetheless be produced with a synthesizer capable of only two sorts of excitation: buzz and noise. For phoneticians as much as for the phonologists they often criticize, it is a convenient if inaccurate simplification to describe speech sounds as either voiced or voiceless. The problem represented by the voiced/voiceless decision in automatic speech analysis and resynthesis [13] reflects the articulatory truth that the larynx is capable of more than two states [14]. The naturalness of synthetic speech can be improved substantially by allowing the formant filters to be excited by a mixture of buzz and noise excitation [15], and speech synthesizers have increasingly included dynamic glottal source control [16,17, 18].

The experiments have focused on a stretch of speech of barely 800ms duration, leaving unaddressed issues concerned with longer utterances. Alterations in **ps** (subglottal pressure) have greater influence on naturalness over longer stretches of speech, and linguistically sensitive use of **ps** yields notable improvements in naturalness. For example, dividing the copy-synthesized reading of Poe’s *The Raven* (provided with Illsyn) into stressed and

unstressed syllables (*cf.* Section 3.2) and assigning **ps** of 9cm H₂O to the stressed, and 6cm H₂O to the unstressed, creates an impression of rhythm, generating a more natural sounding “reading”.

There are considerable possibilities that can only be briefly mentioned here. Both time-varying parameters (**ap**, area of posterior glottal chink) and speaker constants (open quotient, spectral tilt) could be used to model the effect of different voice qualities on the proportion of devoiced vowels, since certain voice qualities, such as breathy voice, may predispose a speaker to use devoicing strategies in preference to other attested available ones, such as flapping or lenition. The articulatory parameters of HLsyn also allow flexible modelling of further rate-dependent Connected Speech Processes, and synthesizing natural-sounding longer stretches of speech would give insights into the effects of subglottal pressure and changes in vocal tract and vocal fold compliance of which the short tokens synthesized here give a glimpse.

NOTES

¹By convention, **bold lower-case text** refers to **HLsyn parameters** and **CAPITAL LETTERS** to **KLATT PARAMETERS**.

REFERENCES

- [1] Rodgers, J. (1998). *Vowel devoicing in English*. Ph.D. thesis, University of Cambridge.
- [2] Rodgers, J. (1999). Segmental and suprasegmental influences on the realization of voicing in English vowels. In J. Local and R. Ogden (Eds.), *Papers in Laboratory Phonology VI: constraints on phonetic interpretation: phonetic interpretation and its relation to linguistic systems*. Cambridge: Cambridge University Press.
- [3] Sensimetrics Corporation (1995a). Notes on HL synthesis. *HLsyn manual*.
- [4] Bickley, C., K. Stevens, and D. Williams (1997). A framework for synthesis of segments based on pseudoarticulatory parameters. In J. van Santen, R. Sproat, J. Olive, and J. Hirschberg (Eds.), *Progress in speech synthesis*, 211–220. New York: Springer.
- [5] Klatt, D. (1975). Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of Speech and Hearing Research* 18, 686–706.
- [6] Stetson, R. (1951). *Motor Phonetics (A Retrospective Edition)*, J. Kelso and K. Munhall (Eds.). Boston: College-Hill.
- [7] van den Berg, J. (1968). Mechanism of the larynx and the laryngeal vibrations. In B. Malmberg (Ed.), *Manual of phonetics*, 278–308. Amsterdam: North-Holland.
- [8] Stevens, K. (1977). Physics of laryngeal behavior and larynx modes. *Phonetica* 34, 264–279.
- [9] Shadle, C. (1997). The aerodynamics of speech. In W. Hardcastle and J. Laver (Eds.), *The handbook of phonetic sciences*, 33–64. Oxford: Blackwell.
- [10] Svirsky, M., K. Stevens, M. Matthies, J. Manzella, J. Perkell, and R. Wilhelms-Tricarico (1997). Tongue-surface displacement during bilabial stops. *Journal of the Acoustical Society of America* 102, 562–571.
- [11] Sensimetrics Corporation (1995b). Exercises from summer course at Mariefred, Sweden. *HLsyn manual*.
- [12] Hawkins, S. and A. Slater (1994). Spread of CV and V-to-V coarticulation in British English: implications for the intelligibility of synthetic speech. *Proceedings of the International Conference of Spoken Language Processing (ICSLP) II*, 57–60.
- [13] Darwin, C. and M. Pearson (1982). What tells us when voicing has started? *Speech Communication* 1, 29–44.
- [14] Pierrehumbert, J. and D. Talkin (1991). Lenition of /h/ and glottal stop. In G. Docherty and R. Ladd (Eds.), *Papers in Laboratory Phonology II: gesture, segment, prosody*, 90–117. Cambridge: Cambridge University Press.
- [15] Makhoul, J., R. Viswanathan, R. Schwartz, and A. Huggins (1978). A mixed source model for speech compression and synthesis. *Journal of the Acoustical Society of America* 64, 1577–1581.
- [16] Fant, G. and T. Ananthapadmanabha (1982). Truncation and super-position. *Speech Transmission Laboratory Quarterly Progress Status Report, Royal Institute of Technology, Stockholm* 2–3, 1–17.
- [17] Klatt, D. and L. Klatt (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America* 87, 820–857.
- [18] Stevens, K. (1991). The contribution of speech synthesis to phonetics: Dennis Klatt’s legacy. *Proceedings of the XIIIth International Congress of Phonetic Sciences I*, 28–37.