Modelling Voiceless Speech Segments by Means of an Additive Procedure Based on the Computation of Formant Sinusoids

Ingo Hertrich, Hermann Ackermann

Department of General Neurology, Hertie Institute for Clinical Brain Research, University of Tübingen, Germany

ingo.hertrich@uni-tuebingen.de

Abstract

A previously developed vowel synthesis algorithm implements formants as sinusoids, amplitude- and phase-modulated by the fundamental frequency (Hertrich and Ackermann, 1999, Journal of the Acoustical Society of America, 106, 2988-2990). The present study extends this approach to the modelling of the acoustic characteristics of aperiodic speech segments. To these ends, a voiceless signal component is generated by adding at each sample point a random parameter onto the formants' phase progression. Voiceless stop consonants then can be modelled, e.g., by combining a release burst, i.e., an interval in which the formant sinusoids abruptly increase and gradually decrease in amplitude, with formant-shaped noise components, representing inter-articulator frication, aspiration, and breathy vowel onset.

1. Introduction

Most speech synthesizers use a source-filter model in order to generate acoustic output signals. As a rule, the vocal tract filter characteristics are either derived from articulation-based parameters or specified in terms of a formant structure while the source, in case of voiced speech segments, approximates the laryngeal excitation signal [1]. A different approach is used in sinusoidal coding techniques [2, 3, 4, 5]: Each fundamental period, i.e., the time domain of a single laryngeal cycle, can be approximated by summing up a set of partial "formant-wave-functions" [3], corresponding to the eigenfrequencies of the vocal tract during this period. At least some of the remaining aspects of the speech signal such as formant bandwidth and voice quality can be implemented as modulations of the amplitude envelopes of these waveforms, characterized by, e.g., by an initial attack interval followed by a decay function within each fundamental period. Thus, formants can be modeled as sinusoids, implementing fundamental frequency as an amplitude- and phase modulation of the formants. In comparison to the natural mechanisms of speech production as well as most speech synthesizers, this approach reverses the source-filter hierarchy and, thus, might be considered an artificial construct. However, amplitude-, phase- and frequency-modulated sinusoids provide the opportunity for a more explicit control of formant structure at a high temporal resolution and at a high computational precision and, therefore, can be used to produce well-defined speech-like stimuli for the purpose of listening experiments. For example, a recent magnetoencephalographic study on the auditory processing of voiced stop consonants used this method to create stimuli that exclusively differed in the duration of syllable-initial formant transitions [6]. As a further example, an investigation of dichotic listening effects compared the perception of natural /ba/ and /da/ syllables to synthetic cognates that exclusively differed in their syllable-initial formant transitions [7]. Apart from psychoacoustic research, formant waveform synthesis may also contribute to an extension of speech synthesis applications with respect to some dynamic aspects of speech such as the continuously changing formant structure, in stop consonant release transients [8], characterized by a time-varying formant structure following a single excitation burst.

As compared to vowels, consonants may exhibit a more complex vocal tract resonance structure due to the engagement of different sound sources and a compartmentalization, more or less, of the vocal tract [9]. While resonance functions following impulse-like excitations can easily be created by sinusoidal formant wave functions, the formant-based generation of voiceless segments seems to be more difficult. One possibility of handling aperiodic segments is the use of multiple overlapping formant waves with irregularly-timed temporal onset [10]. Alternatively, voiceless formant waveforms may be derived from continuous narrowband-sinusoids lacking any decay function. This approach gives rise to unnaturally-sounding whistling-like sounds that, however, still can be perceived as speech ("sinusoidal speech") under some circumstances [11]. In order to simulate a noise-like source, these sinusoids can be manipulated by adding a random factor on their sample-to-sample phase progression [12], resulting in an increase of bandwidth. In fact, this procedure manipulates the instantaneous frequency of the formant waveform while the amplitude and the center frequency can be kept constant. In fact, fricative consonants such as /s/ are often characterized by high and sharp resonance frequencies due to the presence of small cavities near the sound source, giving rise to whistling-like phenomena [13], which can easily be modelled by this kind of synthesis.

The present study represents an extension of the formant-waveform-based speech synthesis algorithm introduced by Hertrich and Ackermann [5]. 'Voiceless' signal components are realized by random phase perturbation of the formant waves. If the perturbation is set to zero, the formants of the aperiodic signal component correspond to pure sine waves. In case of small perturbations, the formants remain visible in the spectrogram, and spectral density near the formant frequencies is relatively high. Increasing the random component ultimately results in broadband noise.

2. The Algorithm

The acoustic target characteristics of the signal to be synthesized are specified in an ASCII input file. Each line of this text file contains a set of 18 parameters referring to duration, intensity, voicing characteristics, fundamental frequency (F0), and five formant frequencies (F1 to F5) as well as the relative formant amplitudes.

Segment duration (L) represents a time interval of linear changes with respect to the remaining parameters from the current input line to the respective parameter values of the following line (the duration parameter of the final line is not evaluated). These parameters include: voiced (A_v) and voiceless (A_n) signal amplitude, relative amount of phase distortion per sample point for the voiceless formant sinusoids (P_n), relative duration of the rise (V_r) and the stationary phase (V_s) of the amplitude profile during one pitch period of the voiced signal component, fundamental frequency (F0) and its relative amplitude (a_{F0}), and five formant frequencies (F1 - F5) as well as their relative amplitudes in percent of total signal amplitude (a_{F1} - a_{F5}).

As a first step, as in [5], signal portions are also synthesized as sequences of pitch periods, voiced signal amplitude being set to a low value or zero, and the formant-related parameters are interpolated with respect to their values at the begin and the end of the respective pitch period. The second step performs period-by-period synthesis of the acoustic signal according to these specifications.

With respect to the voiced signal component, the algorithm introduced in [5] has been modified in order to provide the possibility to handle the amplitude profile A(t) of the formants within single pitch periods in a more flexible way. To these ends, the formants' amplitude envelope within each pitch period is subdivided into a (linear) rising interval, a steady-state portion, and a final decay phase toward zero at the end of the respective pitch period (t represents time from beginning of the current pitch period):

$$A(t) = A_v \cdot \frac{t}{To \cdot V_r} \qquad \text{for the rise phase,}$$

$$A(t) = A_v \qquad \text{for the steady-state phase, and}$$

$$A(t) = A_v \cdot \frac{To - t}{To \cdot (1 - V_r - V_s)} \qquad \text{for the decay phase.}$$

The voiced signal component of each pitch period, then, is the sum of all formant sinusoids modulated in the following way:

$$y_v(t) = A(t) \cdot \left[a_{F0} \cdot \sin\left(2\pi \cdot \frac{t}{To} \right) + \sum_{i=1}^{5} a_{Fi} \cdot \sin \varphi_{Fi}(t) \right],$$

where a_{F0} is the relative amplitude of the fundamental frequency, a_{Fi} is the relative amplitude of the i^{th} formant, and the phase angles

$$\varphi_{Fi}(t) = \varphi_{Fi}(t - dt) + 2\pi \cdot F_i(t) \cdot dt$$

are computed incrementally [$\varphi_{Fi}(0) = 0$] for each sample point using the instantaneous formant frequencies

$$F_i(t) = F_i(0) + \frac{t}{To} \cdot \left[F_i(To) - F_i(0) \right]$$

(dt is the duration between two successive signal samples).

In contrast to the voiced part, the voiceless signal component does neither exhibit pitch-induced amplitude modulation nor a phase reset at the beginning of each pitch period. Each formant is just represented by a sinusoid of a given amplitude

$$y_n(t) = A_n(t) \cdot \sum_{i=1}^{5} a_{Fi} \cdot \sin \varphi_{Fi}(t),$$

where the formants' phase angles are derived by, first, considering (as in the voiced signal component) the instantaneous formant frequencies $F_i(t)$ and, additionally, a random increment the magnitude of which may also vary in time:

$$\varphi_{Fi}(t) = \varphi_{Fi}(t - dt) + 2\pi \cdot F_i(t) \cdot dt + P_n(t) \cdot rnd,$$

rnd being a random number in the range +/- 2π and

$$P_n(t) = P_n(0) + \frac{t}{To} \cdot \left[P_n(To) - P_n(0) \right]$$

representing the local value of the phase distortion parameter. For each sample point, then, voiced and unvoiced signal amplitude are added:

$$y(t) = y_v(t) + y_n(t).$$

3. Examples and Comments

3.1. Single-formant Test Signal

In order to demonstrate the working principle of the algorithm, a single-formant test signal was generated, characterized by five phases (0.2 s each) with the following characteristics (Figure 1):

(1) The signal starts with a voiced interval, the formant moving from 500 to 1500 Hz, F0 changing from 250 to 100 Hz, and amplitude decreasing to zero at 0.2 s.

(2) A voiceless segment starts with gradually increasing amplitude, constant formant frequency at 1500 Hz, and

constant amount of phase distortion, resulting in spectral dispersion.
(3) The formant starts moving back to 500 Hz, phase distortion and amplitude being kept constant.
(4) The phase distortion decreases to zero, i.e., the formant approximates a pure sinusoid at 0.8 s.
(5) The pure formant sinusoid rises from 500 to 1500 Hz.

These five phases (segments of a duration of 80 ms each) are exemplified in Figure 2.

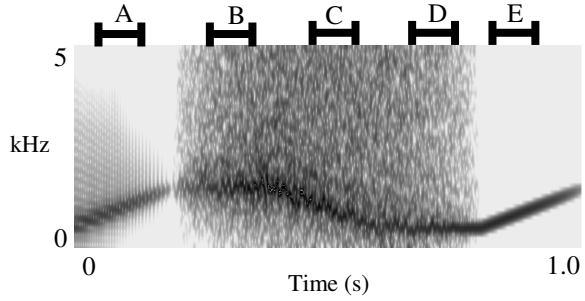

Figure 1. Spectrogram of a single-formant test signal, generated to demonstrate the working principles of the synthesis algorithm. The waveforms of the five 80-ms segments marked with capital letters are displayed in Figure 2.

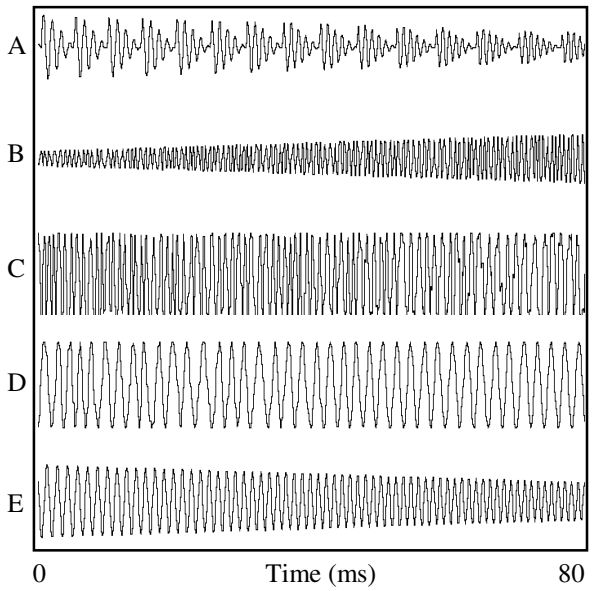

Figure 2. Oscillograms of five selected 80-ms intervals of the test signal displayed as a spectrogram in Figure 1 (capital letters on top of this figure).
A) This segment shows a combination of falling pitch (increase of the fundamental period from left to right), rising formant frequency (oscillations within each period), and decrease in amplitude.
B) Voiceless segment with increasing amplitude, constant effective formant frequency and irregularly-timed peaks indicating the random variation of the formant's instantaneous frequency.
C) Voiceless segment with constant amplitude and increasing period duration (i.e., falling formant frequency. Note that, in contrast to a pure sinusoid, the shape of the waveform and the spacing of peak-to-peak intervals is characterized by some irregularity.
D) During this segment, the random factor upon the formant's instantaneous frequency continuously decreases, resulting in increasingly regular peak-to-peak intervals from left to right.
E) Sinusoid with decreasing amplitude and rising frequency.

3.2. Synthesis of a voiceless stop consonant-vowel syllable

In principle, aspirated stops may encompass five acoustic events, (1) a silent occlusion interval, (2) the initial plosion burst, (3) a short interval of inter-articulator frication, (4) aspiration noise, and (5) vowel onset. To some degree, these intervals may overlap, eventually giving rise to spirantization, multiple initial bursts, inter-articulator frication synchronous with aspiration noise, and/or a breathy or harsh voice quality during the initial part of the vowel. As an example, the syllable /ka/ (Figure 3) was synthesized using the parameter specifications listed in Table 1. The relevant phonetic characteristics of this signal were composed in the following way: The initial plosion and the amplitude decrease following the burst is represented in the first interval (20 ms, lines 1-2 in Table 1), modeled as a superimposition of phase-distorted (A_n, P_n) formant sinusoids and undistorted formant sinusoids declining in amplitude (A_v,V_r, V_s). The latter component represents an impulse-like event followed by a vocal tract filter response (see Repp and Lin, 1989) and, thus, can be realized using the algorithm for the synthesis of voiced portions of speech (F0 was set to 50 Hz to obtain a single resonance period within the 20 ms only). The second interval specified in Table 1 (30 ms) mainly models the aspiration phase, exhibiting aperiodic noise with a lower center of gravity than the stop burst. The first formant increases in amplitude (a_{F1}, lines 2-3) while the higher formants undergo attenuation. Note that the first three formant frequencies show a continuous transition typical for velar articulation preceding the vowel /a/ during the initial two intervals (rising F1 and F3, falling F2). The following part of Table 1 (lines 3-6) specifies the vowel part of the syllable, characterized by high voiced signal amplitude (A_v), pitch (120 Hz at vowel onset, declining to 90 Hz), formant frequencies of the vowel /a/, and the largest relative amplitude in the first formant. The initial part of the vowel (lines 3-4) is characterized by an increase of voiced (A_v) and a decrease of voiceless (A_n) amplitude, accounting for declining breathiness during vowel onset. Furthermore, the relative amplitude of the fundamental frequency decreases while the formants are amplified toward the center of the vowel. The offset of the vowel (lines 5-6) shows a drop in voiced intensity, a slight onset of voicelessness, a drop in amplitude of the higher formants, and a change toward a less skewed intensity profile (V_r) within each pitch period.

Table 1. Input parameters used for synthesis of the syllable /ka/ displayed in Figure 3

L (ms)	A_v	A_n	P_n	V_r	V_s	F0 (Hz)	a_{F0}	F1	a_{F1}	F2	a_{F2}	F3	a_{F3}	F4	a_{F4}	F5	a_{F5}
20	5000	2000	.1	.05	.05	50	0	300	.1	1800	.1	1800	.15	3800	.1	4500	.08
35	0	3000	.15	.1	.2	29	.1	500	.1	1600	.15	2000	.12	3800	.08	4500	.05
40	5000	2000	.05	.04	.2	120	.2	800	.15	1240	.1	2300	.05	3800	.05	4500	.03
80	20000	500	.05	.04	.2	110	.1	800	.25	1240	.15	2300	.15	3800	.10	4500	.05
25	20000	0	.05	.04	.2	100	.1	800	.25	1240	.15	2300	.15	3800	.10	4500	.05
0	10000	300	.05	.2	.2	90	.1	800	.25	1240	.15	2300	.05	3800	.05	4500	.03

Abbreviations: A_v = voiced amplitude, A_n = unvoiced amplitude, P_n = phase distortion, L = segment duration, V_r = relative rising time within a pitch period, V_s = relativ duration of the steady-state phase of a pitch period, F0 = fundamental frequency, a_{f0} = relative amplitude of the fundamental frequency, F1 to F5 = formant frequencies, a_{F1} to a_{F5} = relative formant amplitudes (see text for further discussion of the various parameters).

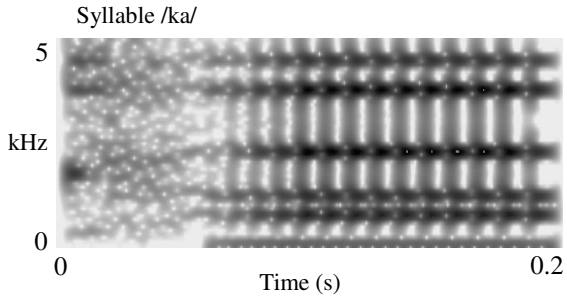

Figure 3. Spectrogram of the syllable /ka/ specified by the parameter settings given in Table 1 (see Text 3.2.). Note the formant transitions in the aspiration phase.

3.3. Comments

The current algorithm allows for the implementation of continuous changes across time of the following aspects of the acoustic signal: (1) voiced and voiceless signal amplitude, (2) formant frequencies and relative formant intensities, (3) magnitude of random phase distortion of the formants controlling spectral bandwidth of voiceless signal components, (4) and the amplitude profile within single pitch periods. So far, this approach does not provide different formant specifications for the voiced and voiceless components in case of mixed voiced/unvoiced signals. Furthermore, the parameter controlling phase distortion during voiceless segments has the same value across all formants. Although the algorithm at its current stage of elaboration seems to work quite well, a more detailed modelling of speech signals may require an increase in the number of input parameters. Considering the additive working principle of this procedure, such extensions can easily be implemented.

4. References

[1] Klatt, D.H., "Review of text-to-speech conversion for English." *J. Acoust. Soc. Amer., Vol. 82, 1987, pp 737-793.*

[2] McAulay, R.J., Quatieri, T.F., "Synthesis based on a sinusoidal representation." *IEEE-ASSP, Vol 34, 1986, pp. 744-754.*

[3] Rodet, X., "Time-domain formant-wavefunction synthesis". *Computer Music J., Vol. 8, 2007, pp. 9-14.*

[4] D'Alessandro, C., "Time-frequency speech transformation based on an elementary waveform representation." *Speech Comm., Vol. 9, 1990, pp. 419-431.*

[5] Hertrich, I., Ackermann, H., "A formant synthesizer based on formant sinusoids modulated by fundamental frequency." *J. Acoust. Soc. Amer., Vol. 106, 1999, pp. 2988-2990.*

[6] Hertrich, I., Mathiak, K., Lutzenberger, W., Ackermann, H., "Processing of dynamic aspects of speech and non-speech stimuli: a whole-head magnetoencephalography study". *Cogn. Brain Res., Vol. 17, 2003, pp. 130-139.*

[7] Hertrich, I., Mathiak, K., Lutzenberger, W., Ackermann, H., "Hemispheric lateralization of the processing of consonant-vowel syllables (formant transitions): effects of stimulus characteristics and attentional demands on evoked magnetic fields." *Neuropsychologia, Vol. 40, 2002, pp. 1902-1917.*

[8] Repp, B., Lin, H.-B., "Acoustic properties and perception of stop consonant release transients." *J. Acoust. Soc. Amer., Vol 85, 1989, pp. 379-396.*

[9] Holmes, J.N., "Formant synthesizers: cascade or parallel?" *Speech Comm., Vol. 2, 1983, pp. 251-273.*

[10] Richard, G., d'Alessandro, C., "Analysis/synthesis and modification of the speech aperiodic component." *Speech Comm., Vol. 19, 1996, pp. 221-244.*

[11] Remez, R.E., Rubin, P.E., Pisoni, D.B., Carrell, T.D., "Speech perception without traditional speech cues." *Sciene, Vol. 212, 1981, pp. 947-950.*

[12] Freed, A., "Spectral line broadening with transform domain additive synthesis." *Proc. Int. Computer Music Conf. Bejing, China, 1999*, http://cnmat.cnmat.berkeley.edu/ICMC99/papers/InverseNoise/InverseNoiseICMC.pdf.

[13] Shosted, R.K., "Just put your lips together and blow? Whistled fricatives in Southern Bantu." Yehia, H.C., Demolin, D., Laboissiere R. (Eds.), *Proceedings of ISSP 2006: 7th Int. Seminar on Speech Production*. CEFALA, Belo Horizonte, 2006, pp. 565-572.