Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

Synthesis strategy

Fant, G. and Mártony, J.

journal: STL-QPSR
volume: 3
number: 2
year: 1962
pages: 020-021

KTH Computer Science and Communication

http://www.speech.kth.se/qpsr

2. Synthesis strategy

The particular combination of source and filter functions are selected in order to make possible the following types of sound segments (compare the section on type features in Chapter IV of ref. [2]).

	Voice source		Noise source		Special requirement
	F	N	F	K	
	Vocalic	Nasal	Vocalic	Fricative	
	A_O	A_N	A_H	A_C	
Phonetic category					
Voiced vowel diphtong, glide [l] and [r] and noise-free variants of [v][w][j]	+				
Voiced occlusive	+				F_1 very low A_O-amplitude low
Whispered vowel			+		
Nasal consonant		+			
Nasalized vowel (Always next to nasal consonant)	+	+			
Fricative consonant including affricates and short initial transient and fricative phase of stop bursts			(+)	+	
Same as above with voicing	+			+	
[h]-like sounds including unvoiced onset of vowel after an unvoiced consonant			+		A_H-source spectrum shaped with low F1-level, low level above 3000 c/s, and -6 dB/oct average slope

The F_1 F_2 and F_3 data are traced directly from spectrograms produced with an expanded time scale (twice the normal Sonagram speed) and expanded frequency scale. The F_0-curve may also be traced in this way from a narrow-band spectrogram following one of the harmonics. Alternatively, we produce an F_0-curve with an automatic pitch extractor

recording it as a Mingogram with the appropriate time scale for transfer to the coding sheet of the function generator.

An A_0-curve is also transferred from a Mingogram. As a crude approximation to the inverse filtering we make use of a simple integration for deriving the source amplitude from the speech wave, this gives us A_0+A_N, a separation is made with the aid of a spectrogram. After the first trial of making a synthesis and comparing it with the original it is possible to make fine adjustments of these gating functions as well as all other functions.

Fricatives of the type s and sh can generally be synthesized with a fairly good quality merely by placing K_1 K_2 and K_0 according to the visible evidence from the spectrogram. For the intense syllables such as [s] and [ʃ] and [ç] the formant K_1 is placed on the main peak, K_2 on the next higher peak of importance, and K_0 about an octave below K_1. In case of labiodental [f] K_1 and K_0 are placed rather close in the 1000-2000 c/s range whereas K_2 is placed at the upper extreme of the frequency scale in the vicinity of 8000 c/s and is well damped. The F-system activated through gate A_H is frequently added if judged necessary by the appearance of F2 F3 and F4 in addition to higher K-type formants. This would be common practice for an aspirated [t]-burst. The A_H and A_C sources are shaped from white noise. The A_C source is flat and A_H is integrated (-6 dB/oct) and both have some degree of high-pass filtering to suppress the spectrum level below 500 c/s. The A_H-gate has in addition a low-pass cutoff at 3000 c/s. Without the latter precaution an [h]-sound might be confused with a [sh]-sound.

The A_0-source has an average slope of -12 dB/oct. The effects of radiation transfer are added by a differentiation (+6 dB/oct) in the A_0 shaping circuitry. Variations of the voice source are made possible by an additional conjugate complex zero and a pole and a few zeros and poles on the negative axis (simple RC circuitry).

Only in very special cases we have bothered to attempt a more detailed source match. In the sentence "I enjoy the simple life", see Fig. II-6, synthesized by J. Holmes ref. (3), the primary voice source was takne from an especially shaped triangular wave selected to

match the voice source pulse shape as viewed from an inverse filtering set up.

The quality of the synthetic speech can be made to approach the human original very closely providing a careful match has been undertaken and the particular speaker has voice characteristics which are favorable for reporduction with OVE II. In general, basing the synthesis on a standard source, we may loose typical aspects of the speaker's voice timbre. However, apparent speaker characteristics are still retained in the faithful reproduction of the F-pattern (F_1 F_2 and F_3) and F_0.

The general impression of the OVE II speech is that it lacks the "harshness" quality typical of most channel vocoders, but that it also fails to reproduce the elements of "crispness" found in many human voices.

3. Quantization of synthesis parameters x)

A quantization scheme planned to suit the demands of a synthesis process in a formant vocoder was carried out by drawing staircase curves instead of continuously varying curves. Fig. II-7 illustrates this procedure which was carried through for two sentences ("I enjoy the simple life" and "He knows just what he wants"). The parameters were sampled at a rate of 40 times per second and were quantized as follows:

Parameters	Number of bits	Number of levels	Comments
F_0	4	16	First voiced sample following voicelessness was coded in 1/6 octave steps covering the range 60-340 c/s. Next and following samples within voiced portions of speech were coded in 1/24 octave steps of change vs the pitch of the previous sample. Total possible range of F_0 46.5-428 c/s.
F_1	4	16	The range of F_1 was 150-900 c/s covered in 50 c/s quantal steps.
F_2	4	16	The range of F_2 was 550-2800 c/s covered in 150 c/s quantal steps.

x) Section 3 contains old material, first reported in STL-QPSR 1/1961. It is included with the purpose of making chapter II a complete summary of present synthesis techniques.