ARCHIVES OF ACOUSTICS
31, 4 (Supplement), 177–181 (2006)

SINE-WAVE WINDOWED SYNTHESIS

H. GARDZIELEWSKA, T. KACZMAREK

Adam Mickiewicz University
Institute of Acoustics
Umultowska 85, 60-614 Poznań, Poland
e-mail: hania@spl.ia.amu.edu.pl

(*received June 15, 2006; accepted September 30, 2006*)

Speech can be understood even when a 3-tone replica of speech is presented to a listener. In tonal synthesis, called sine-wave synthesis (SWS), the output signal consists of number of time-varying sinusoids that follow center frequencies and amplitudes of the first number of formants of a natural utterance. In this paper we propose an alternative technique of speech synthesis. It is based on the number of dominant frequency components present in the original signal. In the proposed method, called sine-wave windowed synthesis (SWWS), the amplitudes and frequencies of tonal components are changed in discrete steps in subsequent time windows. Perceptual tests performed on Polish speech show that signals synthesized with the SWWS technique are judged as more natural and intelligible than SWS speech.

Key words: sinewave synthesis, Polish speech intelligibility, speech coding.

1. Introduction

Speech coding techniques attempt to minimize the bit rate in the digital representation of a signal while maintaining the phonetic information and the required signal quality, meaning its naturalness and diversity for each talker. Speech coding techniques can be divided into waveform coding and spectral coding. The waveform coders try to reproduce the speech waveform as faithfully as possible, and are able to produce high quality-speech but at rather high bit rates. The spectral coding techniques are based on the spectral feature of speech. Speech synthesized with these techniques characterize the lower levels of speech quality at a low number of bits used for synthesis. Sinewave synthesis (SWS) can be an example of spectral synthesis. SWS is based on linguistic information, resistance to distortion and spectral information reduction, which has been confirmed by numerous research results, obtained among others by REMEZ *et al.* [4, 5]. In SWS the speech spectrum is coded in the number of time-varying sinusoids equal in frequency and amplitude to the respective peaks of the first successive formants of a natural-speech utterance. Regardless of the impressive information reduction the linguistic information is to a large extent preserved. Most of the studies on intelligibility

of SWS compressed sounds (i.e. REMEZ *et al*., [5]; MCAULAY, [3]) refer only to English, which is a vowel-dominated language. In contrast to English, the Polish language is consonant-dominant. The key objective of the present study was to analyze and select from three proposed synthesis methods the best one for facilitating the highest level of Polish speech intelligibility at a low compression level. The tested methods were: two modified versions of sinewave synthesis elaborated in Adam Mickiewicz University and the original method elaborated at Haskins Laboratories.

2. Sinewave speech synthesis method: Overview

2.1. SWS

In the original version of SineWave Synthesis, called here method A, sinewave speech parameters were extracted using LPC analysis. A naturally pronounced utterance was cut off above 8 kHz within the routine to focus the LPC on the main formant region below 4 kHz. The output signal consisted of a number of time-varying sinusoids that followed the LPC-derived center frequencies and amplitudes of the first number of formants of a natural utterance.

2.2. Modification of SWS

Two modified methods, B and C were proposed. In these methods, a cepstrum analysis was used instead of the LPC algorithm. The frequency components with the highest amplitude were extracted, instead of the exact formant frequencies. The amplitudes and frequencies of the dominant frequency components were derived with 20 ms resolution.

Because of the large amount of energy in the high frequency range in Polish speech (JASSEM, [2]), the range of dominant frequency components tracking incorporated a band from 200 Hz up to 8 kHz, at a sampling frequency equal to 16 kHz.

In method B, as in the original SWS method, the dominant frequency components' time pattern were interpolated and the amplitude and frequency of each dominant frequency component was calculated for each sample of the reconstructed signal. The synthesized speech consisted of time-varying continuous sinusoidal patterns that resembled the computed peaks of the changing dominant frequency components of the natural utterance.

In method C, called further sinewave windowed synthesis (SWWS), the signal was synthesized in 40 ms time windows. The length of the Hanning window used in this method corresponded to the double length of the resolution of the analysis algorithm. It resulted in the 50% overlapping-required for smooth reproduction. Each window contained a number of sinusoids corresponding to the number of dominant frequency components calculated earlier. The frequencies and amplitudes of all sinusoids within the window were constant and were taken from analysis data. The synthesized speech consisted of window-varying discrete sinusoidal structure.

3. Experiment

3.1. Subjects

30 participants aged 20 to 24, took part in the experiment. All of them were students of Adam Mickiewicz University's Physics Faculty. The participants had no previous experience in synthetic speech intelligibility assessment and were paid for their participation in the experiment.

3.2. Speech material and equipment

The CORPORA multitalker database, designed for automated recognition of Polish speech (GROCHOLEWSKI, [1]) was used for testing. All the sentences were meaningful, declarative, interrogative or imperative statements such as, "Wór rur żelaznych ważył" ("He weighed a sack of iron tubes"). 108 sentences were picked at random from the database. The sentences were divided into four lists, each with a different compression level. The compression level was 9, 6, 4 and 3 tones. Each list contained 27 sentences. The average number of words in each sentence was 5, so that gave approximately 140 words for each list. The sentences of each list were pronounced by one female speaker. The signals were presented at 65 dB SPL.

3.3. Procedure

Participants were randomly divided into three groups of 10 persons. Each group listened to the sentences generated with the method A, B or C (3 conditions: A, B, C). All participants first listened to signals synthesized with 9 tones and then with 6, 4 and 3 tones (4 conditions: 9, 6, 4, 3). Between consecutive listening sessions there was a break of at least 30 minutes. Participants typed the contents of each utterance the way they heard it in a special dialogue box. The utterance typed by each participant was then compared to the original utterance. The twelve experimental conditions were named as follow: 9A, 9B, 9C, 6A, 6B, 6C, 4A, 4B, 4C, 3A, 3B, 3C. On the basis of the collected results the word's intelligibility expressed in percentage was assessed.

3.4. Results

The methods applied to the same speech material resulted in three different spectra. The main differences among these spectra can be seen in Fig. 1.

In Fig. 2 the intelligibility results are displayed for the 9-, 6-, 4- and 3-tone replica of speech for each of the three synthesis methods examined in this experiment.

It is worth noticing that in the method A, the speech synthesized with 3 tones was almost as intelligible as speech synthesized with 4 or 6 tones. For 9-tones speech was reported as completely unintelligible.

The results of experiment 1 showed that the highest level of speech intelligibility was achieved with method C speech synthesis.

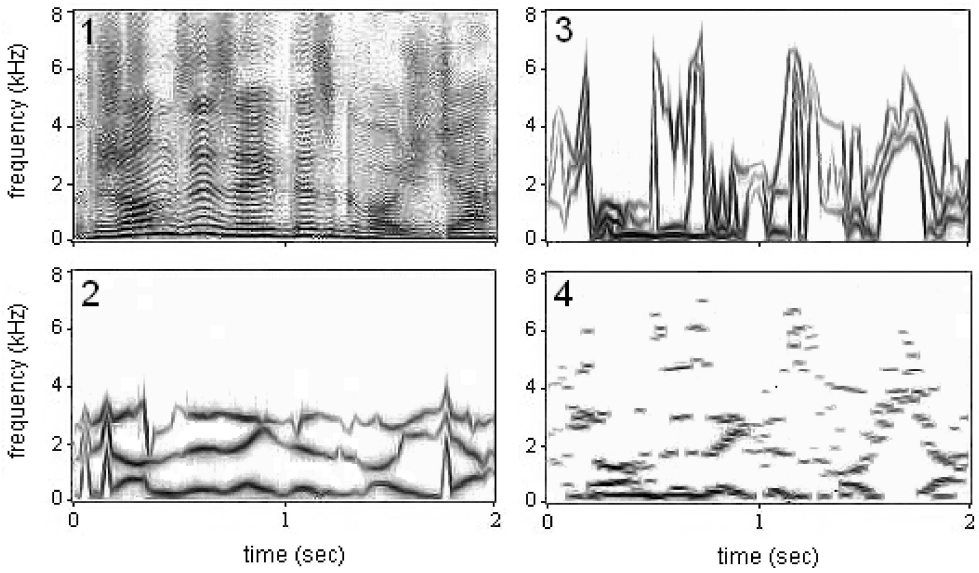

Fig. 1. 1 – Spectrogram of the original utterance "W żądzy zejdę z gwoździa"; 2 – spectrogram of the 3-tone sinusoidal replica synthesized with method A; 3 – spectrogram of the 3-tone sinusoidal replica synthesized with method B; 4 — spectrogram of the 3-tone sinusoidal replica synthesized with method C.

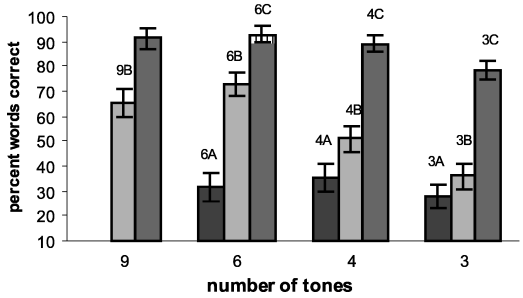

Fig. 2. The averaged percentage of words correct in 2s-long utterances for the number of tones used for synthesis (9, 6, 4, 3) for each of 3 tested methods. Error bars indicate values of a standard deviation.

4. Discussion

Comparison of the intelligibility results for sentences synthesized with the original SWS method and with the SWWS method shows how broadening the speech signal analysis band up to 8 kHz is crucial for proper perception, at least in the case of the Polish language. Taking into account the theory of the simultaneous processing of linguistic and personal information, as well as their common influence on speech intelligibility, we rejected the assumption that first three formants are necessary for the best intelligibility results. This decision was also influenced by the fact that in the case of the Polish language 12 vowels require higher formants of frequency above 3 kHz for their proper identification (JASSEM, [2]).

Attention should also be paid to subjects' reaction to synthesized sounds. Many participants describe the sounds synthesized with the SWS as made of impulses and squeaks, unpleasant in audition. Signals synthesized with the SWWS were described as sounding more natural.

References

[1] GROCHOLEWSKI S., *CORPORA-Speech Database for Polish Diphones*, Proc. Eurospeech'97, pp. 1735–1738 (1997).

[2] JASSEM W., *Basis of acoustical phonetic* [in Polish], chap. 13, pp. 218–230, Polish Scientific Publishers PWN, Warszawa 1973.

[3] MCAULAY R. Q., QUATIERI T. F., *Speech analysis/synthesis based on a sinusoidal representation*, IEEE Trans. ASSP, **34**, 744–754 (1986).

[4] REMEZ R. E., RUBIN P. E., BERNS S. M., PARDO J. S., *On the perceptual organization of speech*, Psychol. Rev., **101**, 129–156 (1994).

[5] REMEZ R. E., RUBIN P. E., PISONI D. B., CARREL T. D., *Speech perception without traditional speech cues*, Science, **212**, 947–950 (1981).