ARCHIVES OF ACOUSTICS
**32**, 3, 579–589 (2007)

# THE INTELLIGIBILITY OF POLISH SPEECH SYNTHESIZED WITH A NEW SINEWAVE SYNTHESIS METHOD

Hanna GARDZIELEWSKA, Anna PREIS

Adam Mickiewicz University
Institute of Acoustics
Umultowska 85, 61-114 Poznań, Poland
e-mail: hania@spl.ia.amu.edu.pl, apraton@amu.edu.pl

(*received May 09, 2007; accepted July 25, 2007*)

SineWave Synthesis, (SWS), allows a significant reduction of the information carried by a speech signal representing by the dynamic spectral properties of formants selected from the natural speech. The synthesis rejects all the detailed acoustic information carried by a signal, including the fundamental frequency as well as harmonic and noise components. Regardless of the impressive information reduction (the compression coefficient for 3-tone synthesis reaches even 195:1), the linguistic and extra-linguistic information of a signal are to a large extend preserved. For the first time, a modified version of SWS was used to analyze Polish speech in order to evaluate the relationship between data reduction and the intelligibility of speech. Speech intelligibility was tested in different utterances varying in grammatical structure, linguistic information, and duration. The modified SWS method, elaborated in Adam Mickiewicz University in Poznań, provided noticeably better results for Polish speech than the original method elaborated in late 1970s at Haskins Laboratories.

**Keywords:** sinewave synthesis, synthetic speech perception, Polish speech intelligibility, linguistic and extralingusitic information

## 1. Introduction

The term "speech signal" is strictly related to the concept of information. Information carried by a speech signal relates to the features of the talker (in the form of extra-linguistic information) and to the message the talker wishes to impart (in the form of linguistic information).

Information carried by a speech signal is very resistant to distortion and spectral information reduction, as has been confirmed by numerous research results obtained, among others, by REMEZ *et al.* [1, 2]. In their studies they treated speech signals with the SineWave Synthesis. In this synthesis, the changing pattern of vocal resonances (formants) is modeled by a limited number of tones reflecting the spectral dynamics and the structure of the signal. The reproduced sounds lose their naturalness but still remain intelligible. Most of the studies on intelligibility of SWS compressed sounds (e.g. [1, 3, 4]) refer only to English, which is a vowel-dominated language. To the best of the authors' knowledge, no detailed analysis on the quality of SWS results in respect to a consonant-dominated language, like Polish, has been performed before.

The key objective of the present study was to determine the relationship between speech data reduction (the number of tones reproducing the speech signal) and the intelligibility of synthetic speech. A more specific aim was to determine how synthetic speech intelligibility depends on the linguistic, especially syntactic and semantic, and extra-linguistic information it carries. Speech intelligibility was tested in different utterances varying in the grammatical and logical structure of their linguistic information, and the acoustic characteristics of the talker. A new sine-wave synthesis method was developed specifically for the analysis of Polish speech.

## 2. Experiment I

### *2.1. Method*

In the original Matlab version of SineWave Synthesis created by S. Frost and P. Rubin, of Haskins Laboratories, based on routines provided by D. Ellis, of the International Computer Science Institute, Berkeley, CA, [5], sinewave speech parameters were extracted using LPC analysis. A naturally pronounced utterance was analyzed in the range up to 4 kHz in 20-ms frames. Amplitude and frequency values of the first successive formants were derived with 20-ms resolution. The formants' time pattern (collected every 20 ms) were interpolated and amplitude and the frequency of each formant was calculated for each sample of the reconstructed signal. The output signal consisted of a number of time-varying sinusoids that followed the LPC-derived center frequencies and amplitudes of the first successive formants of a natural utterance.

In the modified method, the synthesis was based on dominant frequency components. A given number of dominant frequency components were identified after cepstral smoothing of the speech signal. Only the frequency components with the highest amplitudes were reproduced. Because of the large amount of energy in the high frequency range in Polish speech [6], the range of dominant frequency components tracking incorporated a band from 100 Hz (200 Hz for a female speaker) up to 8 kHz, at a sampling frequency of 16 kHz. The signal was synthesized in 40-ms frames. Each frame was multiplied by a 40-ms Hanning's window. The Hanning's window used in this method corresponded to the double length of the analysis algorithm that was equal to 20 ms. It resulted in 50% overlapping. Each window contained a number of sinusoids corresponding to the number of dominant frequency components. 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.

### 2.2. *Subjects*

Forty four participants (10 women and 34 men) aged 20 to 24, took part in the experiment. All of them were students at the Adam Mickiewicz University. The participants were native talkers of Polish who reported no past or present hearing disorders and qualified as having normal hearing (normal hearing was defined as the audiometric threshold 20 dB HL or better, for a frequency range from 250 Hz to 8000 Hz) [7]. The participants had no previous experience in synthetic speech intelligibility assessment and were paid for their participation in the experiment.

### 2.3. *Speech material and equipment*

The CORPORA multitalker database, designed for automated recognition of Polish speech [8] was used for testing. CORPORA contains 114 low-redundancy sentences, each 2 s in duration, pronounced by 37 different talkers. All sentences were meaningful, and were either declarative, interrogative or imperative statements such as *Wór rur żelaznych ważył* or *Żmije są sine bo są zimne* ("He weighed a sack of iron tubes" and "Vipers are blue because they are cold"). Eighty one sentences were picked at random from the database. The sentences were divided into three lists, each with a different compression level. The compression level was 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 either by different talkers or by a single, male or female talker, depending on the choice of experimental conditions. The entire test material was stored in the computer memory and its output was via custom software routines using the MATLAB software package. The stimuli were presented through Sennheiser headphones, inside a specially designed sound attenuated chamber. The signals were presented at 65 dB SPL. The computer was placed outside the chamber during the experimental sessions. Only the computer display and keyboard were allowed inside for the collection of subjects' responses.

### 2.4. *Procedure*

Nine experimental conditions were tested. Participants were randomly divided into two groups of 20 and 24 persons. For each compression level, speech was generated either by different talkers or by single talker (two conditions: DT, ST). The sentences produced by different talkers were presented to the first group and the other group only listened to the sentences generated by a single talker. The last group of participants was further divided into two subgroups: 12 participants listened to utterances presented by a female talker and 12 listened to utterances produced by a male talker (two conditions: STF, STM). This was done in order to take into account the fact that with increasing the fundamental frequency of a voice, the difficulty of determining the formant frequency increases as well. This usually leads to the conclusion that a female voice is less intelligible than a male voice [6, 9]. Comparison of the results obtained from both groups made it possible to evaluate the degree to which a change in the acoustic characteristics of a phonetically distorted signal (extra-linguistic information) affects the perception of linguistic information.

All the participants first listened to signals synthesized with six tones and then with four and three tones (three conditions: 6, 4, 3). There was a pause of at least 30 minutes between consecutive listening sessions.

Participants typed the content 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 nine experimental conditions were named as follows: 6DT, 6STF, 6STM, 4DT, 4STF, 4STM, 3DT, 3STF, 3STM. On the basis of the collected results the word's intelligibility was assessed and expressed as the percentage of correct responses.

### *2.5. Results and discussion*

The speech intelligibility results, expressed as the percentage of average words correct for each list, are presented in Fig. 1. No statistically significant interaction of participants' gender-responses with compression level was obtained. The results obtained from the participants of different gender for each compression level were averaged.

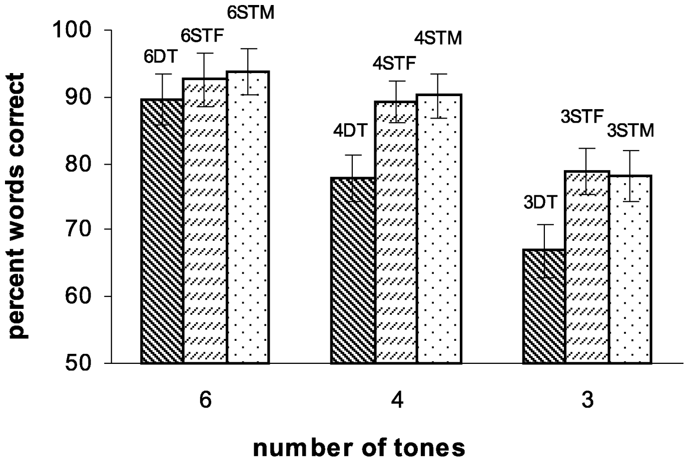

Fig. 1. The averaged percent words correct in 2-s utterances for different numbers of tones used for synthesis (6, 4, 3) for various talker condition (male, STM; female, STF; different, DT). Error bars indicate values of a standard deviation.

A two-way ANOVA showed a significant interaction between the number of tones used for synthesis and the acoustic characteristics of the talkers (DT, STM, STF) $[F(4,123) = 6.69; p < 0.05]$. The Scheffe *post hoc* test showed that the synthetic speech intelligibility results for single talker (both man and female) conditions for 6-, 4- and 3-tones used for speech synthesis did not differ significantly. There were no statistically significant differences between single and different talkers conditions for the 6-tones used for synthesis. In the case of the 3- and 4-tones used for synthesis, different talker conditions had a greater negative impact on the speech intelligibility results when compared to single talker conditions.

The more the speech was compressed, the more talker variability influenced speech intelligibility. Significant differences were observed between single and different-talker conditions for 3- and 4-tone compression levels. The synthetic speech intelligibility results obtained for different talker conditions for 4-tones were equal to the results obtained for single talker conditions for the 3-tones used for synthesis.

The results obtained once more confirm that paying attention to spoken words involves paying attention to voice, which is reflected in the speech intelligibility scores [10]. Despite being prepared to receive a semantic message in such unnatural acoustic conditions, listeners showed evidence of the integral processing of changes in their acoustic environment, namely talker-specific attributes, along with the processing (recognition) of linguistic attributes of a signal [11–13].

## 3. Experiment II

### *3.1. Method*

The synthesis method used in this Experiment was the same as in Experiment I.

### *3.2. Subjects*

Twenty four participants took part in listening sessions in Experiment II (four women, 20 men) aged between 20 and 25. The listeners were already familiar with the synthetic signals – they had participated in the previous synthetic speech intelligibility assessment.

### *3.3. Speech materials and equipment*

The signals presented in this experiment were words selected from a recorded, frequency and phonetics balanced wordlist for the Polish language, elaborated by JASSEM [14]. The signals were divided into two groups. In both groups, the signals were compressed at the level of 3 tones. The first group of signals was composed of 27 utterances, each comprising three words, i.e., *zmierzch, deszcz, gęś* (twilight, rain, goose). The duration of each utterance corresponded to the duration of sentences presented in Experiment I, which was about 2 seconds. The second group of signals was composed of the same set of 81 words, but presented individually. All utterances were generated only by one female talker. Other stimuli presentation details were the same as in Experiment I.

### *3.4. Procedure*

Two experimental conditions were tested. The participants were divided into two groups of 12 (two women and 10 men in each group). One group assessed the intelligibility of 27 utterances built of three unrelated words (3N3 condition). The other group assessed the intelligibility of words presented individually (3N1 condition). Participants typed the contents of each utterance the way they heard it in a special dialogue box. The content of the utterance typed by each participant was compared to the original utterance. On the basis of the collected data the percentage word intelligibility ratio was assessed.

### *3.5. Results and discussion*

In Table 1 the speech intelligibility results are displayed for one- and three-word sentence conditions. The results show no significant effect of participants' gender $[F(1, 20) = 0.32, p > 0.57]$. In the final analysis, the results obtained from the participants of different gender were averaged.

**Table 1.** The averaged percent words correct in three- and one-word utterances for three-tone synthesis.

| | Words correct [%] | Standard deviation [%] |
|---|---|---|
| 3N1 | 85.6 | 4.0 |
| 3N3 | 72.3 | 6.1 |

A two-way ANOVA was conducted. The factors in the design were sentence length (3N3 and 3N1 utterances) and the participants' gender. A highly significant effect was observed for sentence length $[F(1, 20) = 23.28, p < 0.0001]$. The best word-intelligibility assessment was achieved for words presented individually.

The intelligibility of utterances consisting of 3 unrelated nouns decreased by 13%. Three causes responsible for such results may be identified. First, the utterances comprising individual words were three times shorter. Secondly, there were no time gaps between the words in the utterances consisting of three unrelated words – similarly to the real sentences. The words in such utterances were perceived as being produced at a faster speaking rate than they were (compared to the single words). Thirdly, as the nouns were completely unrelated, it was hard for the participants to find a common, logical unity which would help them memorize the words by mutual association. In this case a subject's ability to memorize the presented sequence of the words was an additional factor which influenced the results obtained. This explains the highest standard deviation, 6.1%, obtained in this case.

## 4. Experiment III

### 4.1. *Method*

The synthesis method used in this Experiment was the same as in Experiment I.

### 4.2. *Subjects*

Thirty six persons participated in Experiment III as listeners (six women, 30 men) aged between 20 and 25. The listeners had participated in at least one of the earlier listening sessions and thus were familiar with the synthetic signals.

### 4.3. *Speech materials and equipment*

The signals presented in this experiment were logatoms, nonsense words selected from a structurally and phonetically balanced list of logatoms for the Polish language [15]. In total 243 logatoms were randomly selected. The speech material was divided into two groups. The first one was composed of 81 expressions, each consisting of three logatoms (i.e., *stłopka, czepło, weker*). The expressions (each of 2-s duration) were divided into tree equinumerous lists, each with a different compression level. The compression level was: 6, 4 and 3 tones. The second group of signals was composed of 50 single logatoms compressed at the level of 3 tones, presented individually.

Taking into account the difficulty participants experience when recognizing logatoms, and in order to determine in the most reliable way that synthesis influences their recognition, the original (uncompressed) logatoms were also presented. In the case of signals presented in Experiments I and II, the reference procedure was not necessary, as all uncompressed signals were 100% intelligible. In Experiment III all expressions were generated only by one male-talker. Other stimuli presentation details were the same as in Experiment I.

### 4.4. *Procedure*

Four experimental conditions were tested. The participants were randomly divided into two groups of 19 and 17 persons. The first group assessed the recognition of three logatoms synthesized with six tones and then in turn with 4- and 3-tones (three conditions: 6L3, 4L3, 3L3). After each listening session there was a pause of at least 30 minutes. When they had finished listening to the compressed signals, listeners also listened to the original lists of logatoms. The second group of participants assessed the recognition of 50 single logatoms compressed at the level of 3 tones (one condition: 3L1). After listening to the compressed signals, listeners listened to the same, but uncompressed logatoms. Participants typed the contents of each utterance the way they heard it in a special dialogue box. The content of the expression typed by each participant was compared to the original expression. In total, four conditions were tested: 6L3, 4L3, 3L3, 3L1. On the basis of the collected results, the percentage of correct logatoms recognition was assessed.

### *4.5. Results and discussion*

The recognition of original, uncompressed logatoms was on average 86%, with a standard deviation of 5%, for the three-logatom utterance list and 94% with a standard deviation of 3.1%, for the one-logatom list. In order to reliably display the results obtained in Experiment III, the relative recognition results obtained for four conditions has been shown in Fig. 2. No statistically significant gender-response from the participants was observed. The results obtained from the participants of different gender were averaged.

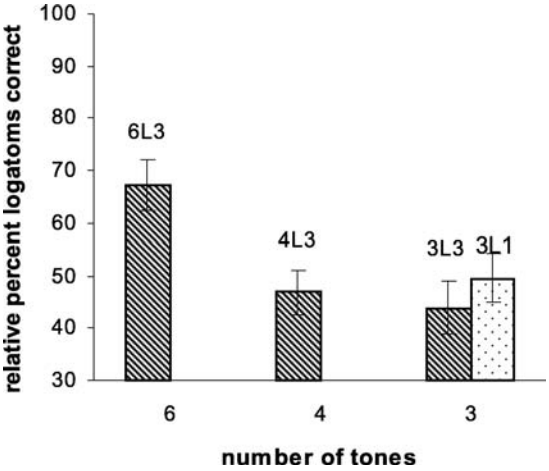

Fig. 2. The averaged relative percent logatoms correct in utterances for different numbers of tones used for synthesis (6, 4, 3). L3 refers to the three-logatom utterances; L1 refer to one-logatom utterances. Error bars indicate values of a standard deviation.

A one-way ANOVA was conducted to assess the data reduction on logatom identification. The analysis was highly significant [$F(2, 54) = 63.59$, $p < 0.05$]. The logatom recognition became poorer as the number of tones used for synthesis decreased. The Scheffe *post hoc* test revealed that the results obtained for 6-tones were significantly different from the results obtained for the 4- and 3-tones used for synthesis. Comparing the above results with the words' intelligibility results from Experiments I and II it was found that in case of words without meaning more information from the sound spectrum must be mapped to achieve comprehensible speech.

In order to estimate the effect of the presented signals' length on logatom identification (3L3, 3L1), statistical analysis was conducted. The results were statistically significant [$F(1, 34) = 20.97$, $p < 0.05$]. One-logatom utterances were easier to recognize than three-logatom utterances. In case of logatoms, apart from missing syntax or logical unity making it easier to memorize a linguistic content, there was an additional recognition hindering factor – the lack of semantic meaning of an expression. The results show how intelligibility/recognition of SWS utterances is strongly based on a person's knowledge and experience, as the expressions heard are fitted to the existing set of words stored in a person's memory [16]. In the case of logatoms there is no such reference.

## 5. General discussion

The results obtained in Experiments I, II and III for 3-tones used for speech synthesis have been shown in Table 2.

**Table 2.** The averaged relative percent words correct in utterances synthesized with 3-tones.

| Speech materials | Utterance duration [s] | Words correct [%] | Standard deviation [%] |
|---|---|---|---|
| Sentences, single talker (ST); Experiment I | 2 | 78.8 | 5.2 |
| Sentences, different talker (DT); Experiment I | 2 | 67.0 | 5.5 |
| Three-word utterances (N3); Experiment II | 2 | 72.3 | 6.1 |
| Three-logatom utterances (L3); Experiment III | 2 | 43.9 | 5.7 |
| Single word (N1); Experiment II | 0.6 | 85.6 | 4.0 |
| Single logatom (L1); Experiment III | 0.6 | 49.5 | 5.3 |

According to the results, the acoustic attributes of a talker (extra-linguistic information) cannot be neglected in speech perception, even in the case of synthetic speech. Diversity of extra-linguistic information has great impact on correct synthetic speech signal identification. The intelligibility of words in sentences uttered by different talkers (Table 2, line 2) was lower by the amount of 12% than the intelligibility of the same words but uttered by single talker (Table 2, line 1).

Utterances devoid of syntactic structure were less intelligible than the equally long logical sentences. Words in utterances devoid of logical and grammatical coherence uttered by a single talker (Table 2, line 3) were better identified by the amount of 6% than words in logical sentences also uttered by single talkers (Table 2, line 1).

The results show that linguistic information reduction through the changes in the sentence length has an even lower impact on speech intelligibility than speaker acoustic characteristic variation. The percentage of words correctly identified in sentences in the single talker condition (Table 2, line 1) was lower by the amount of 7% than in one-word utterances uttered by a single talker (Table 2, line 5).

The results indicate differences in the perceptual processing of words, resulting not only from the physical realization of utterance [17, 18], but also from grammatical and semantic utterance information content. Preservation of grammar and the logical continuity of an utterance significantly facilitate the recognition of individual words.

In the case of logatoms, the lack of any particular meaning made it almost impossible for subjects to reproduce them correctly. The results demonstrate how synthetic speech intelligibility is dependent on the correct perceptual matching of phonetic characteristics of heard sounds with the phonetic characteristics stored in a listener's long-term memory [19–22]. In cases where there were no original phonetic characteristics in a listener's long-term memory, synthetic speech perception turned out to be almost impossible on the basis of such limited acoustic information. Perceptual matching, in principal, facilitates the invariability of a talker acoustic characteristics. The way the speech sounds are generated has a secondary meaning.

## 6. Conclusions

Experiments showed that the intelligibility of synthetic Polish speech depends more on the acoustic characteristics of the talker than on the content of linguistic information. The intelligibility of words in sentences uttered by different talkers was worse by the amount of 12% than utterances of a similar length from one talker. With the characteristics of the talker kept constant, extending the duration of the signal caused intelligibility to deteriorate by 7%, and removing grammatical coherence and cohesion from the utterance caused intelligibility to deteriorate by 6%. The absence of semantic information (i.e. through the employment of logatoms) rendered speech unintelligible.

## Acknowledgments

This research was supported by KBN Grant No. 4T07B 053 26.

## References

[1] Remez R. E., Rubin P. E., Pisoni D. B., Carrell T. D., *Speech perception without traditional speech cues*, Science, **212**, 947–950 (1981).

[2] Remez R. E., Rubin P. E., Berns S. E., Pardo J. S., Lang J. M., *On the perceptual organization of speech*, Psychological Review, **101**, 129–136 (1994).

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

[4] Dorman M., Loizou P., Rainey D., *Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs*, Journal of the Acoustical Society of America, **102**, 2403–2411 (1997).

[5] http://www.ee.columbia.edu/∼dpwe/resources/matlab/sws/.

[6] Jassem W., *Podstawy fonetyki akustycznej*, PWN, Warszawa 1973.

[7] ANSI. ANSI S3.6-1996, *Specifications for Audiometers*, American National Standards Institute, New York 1996.

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

[9] KLATT D. H., KLATT L. C., *Analysis, synthesis, and perception of voice quality variations among female and male talkers*, Journal of the Acoustical Society of America, **87**, 2, 820–857 (1990).

[10] NYGAARD L. C., PISONI D. B., *Speech perception: New directions in research and theory*, [in:] *Speech, Language, and Communication*, MILLER J. L., EIMAS P. D. [Eds.], Academic, San Diego, CA, pp. 63–96, 1995.

[11] MULLENIX J. W., PISONI D. B., *Stimulus variability and processing dependencies in speech perception*, Perception and Psychophysics, **47**, 379–390 (1990).

[12] GREEN K. P., TOMIAK G. R., KUHL P. K., *The encoding of rate and talker information during phonetic perception*, Perception and Psychophysics, **59**, 675–692 (1997).

[13] REMEZ R. E., *Talker identification based on phonetic information*, Journal of Experimental Psychology: Human Perception and Performance, **23**, 651–666 (1997).

[14] JASSEM W., *Frequency and phonetics balanced Polish wordlists*, [in:] *Speech and Language Technology*, JASSEM W., BASZTURA C. [Eds.], Polish Phonetic Association, Poznań, pp. 71–100, 1997.

[15] BRACHMAŃSKI S., STARONIEWICZ P., *Phonetic structure of test material used for subjective speech quality measurements*, [in:] *Speech and Language Technology*, JASSEM W., BASZTURA C. [Eds.], WPN Format, Poznań, pp. 71–80, 1999.

[16] JUSCZYK P. W., LUCE P. A., *Speech perception and spoken word recognition: past and present*, Ear and Hearing, **23**, 2–40 (2002).

[17] REDDY D., *Speech recognition by machine: a review*, Proceedings of IEEE, **64**, 4, 501–531 (1976).

[18] PISONI D. B., LUCE P. A., *Acoustic-Phonetic representations in word recognition*, Cognition, **25**, 21–52 (1987).

[19] MARTIN C. S., MULLENIX J. W., PISONI D. B., SUMMERS W. V., *Effects of talker variability on recall of spoken word lists*, Journal of Experimental Psychology: Learning, Memory, and Cognition, **17**, 152–162 (1989).

[20] JUSCZYK P. W., PISONI D. B., MULLENNIX J., *Some consequences of stimulus variability on speech processing by 2-month-old infants*, Cognition, **43**, 253–291 (1992).

[21] SHANNON R. V., ZENG F. G., WYGONSKI J., KAMATH V., EKELID M., *Speech recognition with primarily temporal cues*, Science, **270**, 303–304 (1995).

[22] MCQUEEN J. M., CUTIER A., NORRIS D., *Flow of information in the spoken word recognition system*, Speech Communication, **41**, 1, 257–270 (2003).