ELSEVIER

Clinical Neurophysiology 110 (1999) 1329–1333

A method for generating natural-sounding speech stimuli for cognitive brain research

Paavo Alku[a,*], Hannu Tiitinen[b], Risto Näätänen[b]

[a]Helsinki University of Technology, Acoustics Laboratory, P.O. Box 3000, FIN-02015 TKK, Helsinki, Finland
[b]Cognitive Brain Research Unit, Department of Psychology, University of Helsinki, Helsinki, Finland

Accepted 16 March 1999

Abstract

Objective: In response to the rapidly increasing interest in using human voice in cognitive brain research, a new method, semisynthetic speech generation (SSG), is presented for generation of speech stimuli.

Methods: The method synthesizes speech stimuli as a combination of purely artificial processes and processes that originate from the natural human speech production mechanism. SSG first estimates the source of speech, the glottal flow, from a natural utterance using an inverse filtering technique. The glottal flow obtained is then used as an excitation to an artificial digital filter that models the formant structure of speech.

Results: SSG is superior to commercial voice synthesizers because it yields speech stimuli of a highly natural quality due to the contribution of the man-originating glottal excitation.

Conclusion: The artificial modelling of the vocal tract enables one to adjust the formant frequencies of the stimuli as desired, thus making SSG suitable for cognitive experiments using speech sounds as stimuli.

Keywords: Speech production; Inverse filtering; Speech synthesis; Speech perception; Auditory discrimination; Mismatch negativity

1. Introduction

Previously, cognitive studies in the auditory domain have mainly been conducted by using non-speech, most often sinusoidal, stimuli (e.g. Rogers et al., 1990; Schröger et al., 1994; Tiitinen et al., 1994). In analysing cerebral processing of speech, there are studies where natural utterances have been used as auditory material (e.g. Zatorre et al., 1992; Kuriki et al., 1995). From the point of view of the naturalness of the auditory stimuli, one should, of course, apply real human speech as stimulus material. Unfortunately, using natural speech is not always possible because the experimental setup typically used in cognitive brain research calls for methods to manipulate the acoustical features (e.g. formants, pitch, duration) of the stimuli. There are many studies where, for example, formants of vowel stimuli need to be adjusted exactly according to a given procedure (Sams et al., 1990; Kraus et al., 1993; Maiste et al., 1995; Sharma et al., 1997). In cases such as those, natural speech cannot be used and the stimulus material is typically produced by commercial speech synthesizers. The quality of the stimuli given by artificial speech synthesis, however, is not equivalent to the quality of natural speech and this might seriously degrade the analysis of the brain activity elicited by the stimuli. In particular, analysing cerebral mechanisms related to language processing calls for synthesis methods that yield stimuli of a highly natural quality. It would be very difficult to analyse reliably, for example, brain activity related to foreign-language learning, or to study speakers of different languages, if the speech stimuli is not of sufficient naturalness.

This paper presents a new method, semisynthetic speech generation (SSG), that was developed in order to synthesize high-quality speech stimuli for cognitive brain research. This new method generates speech signals as a combination of purely artificial processes and those originating from the real human speech production mechanism. The artificial processes of the synthesis approach, on the one hand, enable the full control and manipulation of the formant structure. The man-made processes of the method, on the other hand, bring certain human features (e.g. variation in the length of the fundamental period) into the synthesized speech sound. Consequently, the naturalness of the semisynthesized speech sound is considerably better in comparison to

* Corresponding author. Tel.: + 358-9-451-5680; fax: + 358-9-460-224.
E-mail address: paavo.alku@hut.fi (P. Alku)

1388-2457/99/$ - see front matter CLINPH 98766
PII: S1388-2457(99)00084-X

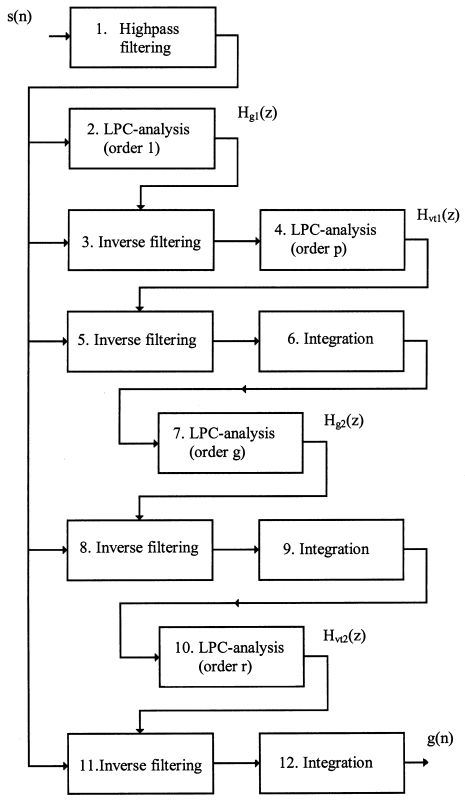

Fig. 1. The block diagram of the IAIF method for estimation of the glottal excitation (g(n)) from the speech signal (s(n)). The model for the vocal tract filtering is determined with an iterative procedure (blocks 2–10) by using LPC-analysis as a computational tool. The estimated glottal flow is obtained by cancelling the effects of vocal tract filtering (block 11) and lip-radiation (block 12) from the speech signal.

conventional speech synthesis that is based on purely artificial processes.

SSG has been used successfully as a method of stimulus generation in several recent studies (e.g. Näätänen et al., 1997; Cheour et al., 1998). The aim of the present paper is to describe, in detail, the structure of this new method and to demonstrate the usefulness of SSG in generating auditory stimuli.

2. Materials and methods

Stimulus generation with SSG comprises two stages. In the first one, the excitation waveform of speech, the glottal waveform, is computed from a real voice signal. In the second stage, the glottal waveform obtained is used as an input to an artificial digital filter in order to produce the desired auditory stimulus. SSG is described below by first presenting how human voice production is modelled. The computation of the glottal excitation that is an essential part of the proposed synthesis method is described next. Finally, the stimulus generation itself with the SSG-method is described.

2.1. Modelling of the speech production mechanism

According to Fant (1960), the production of speech can be modelled by three separate processes: the glottal excitation, the vocal tract filtering, and the lip-radiation effect. The first of these processes, the glottal excitation, corresponds to the pulsating flow of air that comes from the lungs through the vibrating vocal folds. This first process of the human speech production mechanism is named after the orifice between the vocal folds, the glottis. The second process, the vocal tract filtering, corresponds to the strong effect that is created by the physiological filter, the vocal tract, that starts from the vocal folds and ends up at the lips and nostrils. The third process, the lip-radiation effect, corresponds to changing the volume velocity waveform at the lips to a speech pressure signal in a free field at a certain distance from the speaker.

Stimulus generation with the SSG-method is based on the separated speech production model described above. By using this model, it is possible to extract from any given real speech signal the first process, the glottal excitation, which is computed using the inverse filtering technique. By forming an artificial digital model for the other two processes of the model, it is then possible to synthesize a semisynthetic speech signal by using the computed glottal flow as an excitation.

2.2. Computation of the glottal flow with inverse filtering

Inverse filtering is a technique that is used to estimate the source of speech, the glottal flow. An inverse filtering method firstly determines the model for the vocal tract filtering from a given speech signal. It is then possible to cancel the effect of the vocal tract from the speech signal by filtering this through the inverse model of the tract. In SSG, the estimation of the glottal flow is computed using iterative adaptive inverse filtering (IAIF) (Alku, 1992) (see Fig. 1).

As Fig. 1 shows, the only input required in order to estimate the glottal flow with IAIF is the acoustical speech pressure waveform (denoted by s(n) in Fig. 1) captured by a microphone. The output of the method (g(n) in Fig. 1) is the estimated glottal flow (i.e. the first process in the separated speech production model). The method is completely automatic and, if required, it can be implemented to run in real time. As a computational tool for separating different processes of speech production, the IAIF-method uses linear predictive coding (LPC). LPC is a widely applied method in speech processing and its properties are thoroughly described in various textbooks of speech science (e.g. Rabiner and Schafer, 1978). LPC is well suited, especially for the analysis of voice production, because it is able to model the speech spectrum accurately. Hence, LPC can be used as an adaptive method to estimate the vocal tract filtering effect of the separated speech production model. The estimation of the lip-radiation effect is much more straightforward in comparison to the vocal

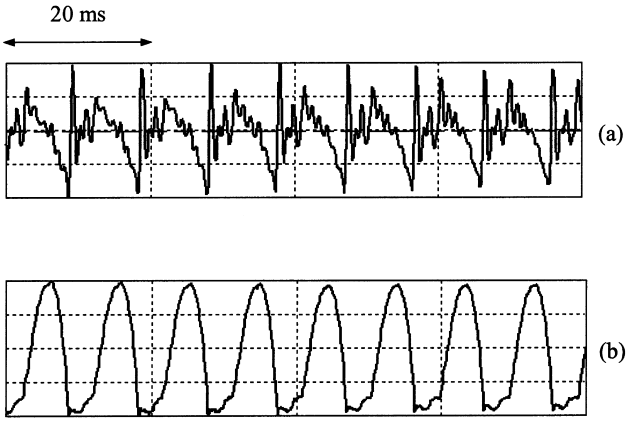

Fig. 2. Speech signal (a) and the estimated glottal flow (b) given by the IAIF method. In generating semisynthetic sounds, the glottal flow is used as an excitation to a digital all-pole filter that produces the desired formant structure. The output of the filter is differentiated according to the lip-radiation effect in order to obtain the semisynthetic speech sound.

tract filtering because the lip-radiation effect can be modelled accurately by a fixed differentiator (Flanagan, 1972). Therefore, cancelling the lip-radiation effect from a signal corresponds to integrating it.

The estimation of the glottal excitation with the IAIF-method consists of the following stages. Firstly (block 1), the speech signal is high-pass filtered in order to remove any distorting low frequency fluctuations captured by the microphone during the recordings. The high-pass filter is a linear phase FIR-filter with a cut-off frequency of 60 Hz. Secondly (block 2), a first order LPC-analysis is computed for the high-pass filtered speech signal. This stage yields a first-order all-zero filter, the transfer function of which is denoted by $H_{g1}(z)$ in Fig. 1, which forms a preliminary estimate for the combined effects of the glottal flow and the lip-radiation effect on the speech spectrum. Thirdly (block 3), the estimated effects of the glottal flow and lip-radiation are cancelled from speech by inverse filtering it through the obtained first order LPC-filter. The output is analysed using a pth-order linear prediction (block 4) in order to obtain a model, denoted by $H_{vt1}(z)$, for the vocal tract filtering. (The order of LPC-analysis, p, is typically between 8 and 12). Next (block 5), the effect of the vocal tract is cancelled from speech by inverse filtering it through the inverse of the obtained pth-order model. A first estimate for the glottal flow is obtained (block 6) by cancelling the effect of the lip-radiation by integrating the output of block 5. The IAIF-method next computes (block 7) a new estimate, denoted by $H_{g2}(z)$, for the contribution of the glottal flow on the speech spectrum by computing LPC-analysis of order g to the obtained first estimate of the glottal excitation. (The value of g is typically between 2 and 4.) By first cancelling the effect of the estimated glottal contribution (block 8) and the lip-radiation effect (block 9), a new model for the vocal tract filtering is obtained by a pth-order LPC-analysis (block 10). The final result is obtained by cancelling the effect of the new vocal tract model (block 11) and the lip-radiation effect (block 12).

An example of a glottal flow estimated by the IAIF-method is shown in Fig. 2b. The speech signal (vowel [a]) from which the analysis was computed is depicted in Fig. 2a. Even though the speech sound was produced using sustained phonation, it can be seen from Fig. 2a that the waveform of the vowel changes from one fundamental period to another. This is a typical feature of natural speech. Unfortunately, this small fluctuation of the speech waveform is very difficult to model properly in conventional speech synthesis. Consequently, the produced speech sound will have machine-like quality. However, it is worth noting that the glottal flow (Fig. 2b) estimated by the IAIF-method also includes small changes from one period to another. Therefore, by using the glottal flow given by IAIF as an excitation in the generation of the stimulus, it is possible to transmit these human-originating features of speech to the synthesized voice signal.

2.3. Stimulus generation based on the separated speech production model

The IAIF-method described in the previous section yields, from any given speech signal, an estimate for the glottal source that has been generated by the vocal folds when the sound was produced. From the point of view of experimental demands of cognitive brain research, an advantageous feature of SSG is that once a glottal waveform has been computed by IAIF it is also possible to synthesize new speech sounds using the same separated speech production model on which IAIF is based. In order to produce different speech sounds, it is necessary that the vocal tract is modelled artificially by a digital filter, the spectrum of which can be adaptively adjusted. Especially in the case of vowel sounds, it is important to be able to create vocal tract filtering effects that correspond to a desired formant structure. This calls for using the digital all-pole filter as a model for the vocal tract. As described by Rabiner and Schafer (1978), the all-pole filter yields a model for the vocal tract transfer function that is accurate, especially for the vowels, but it also provides a good representation for almost all the sounds of speech. The use of all-pole filtering in the modelling of the vocal tract transfer function is justified because the coefficients of the filter can be determined in a straightforward and computationally efficient manner. In particular, it enables transforming the desired formant information from the commonly used F1-F2-space to the coefficients of the digital filter. This can be done by using, for example, the approach presented by Gold and Rabiner (1968), where one formant of the vocal tract is modelled using a second order all-pole filter. The transfer function of this filter can be determined, when the sampling interval, T, of the digitized sound is known, from the frequency, f_i, (in Hz) and the bandwidth, g_i, (in Hz) of the desired formant as follows:

$$H(z) = \frac{1 + r_i^2 - 2r_i \cos b_i T}{1 - 2r_i \cos b_i T z^{-1} + r_i^2 z^{-2}} \quad (1)$$

where $r_i = e^{-2\pi T g_i}$ and $b_i = 2\pi f_i$

The complete digital model for the vocal tract is obtained by cascading several second order filters given by Eq. (1) to synthesize a desired formant structure.

To summarize, the speech stimulus is obtained in the final stage of SSG by first filtering the estimated glottal flow with the digital all-pole filter that has been determined to produce a desired formant structure, and then by differentiating according to the lip-radiation effect, the output of the vocal tract filter.

3. Discussion

3.1. Experiments with semisynthetic stimulus generation

The SSG-technique has been successfully used as a method of stimulus generation in various experiments where auditory information processing of the brain has been studied. Following, we describe as examples some of our recent studies that used the SSG-method in stimulus generation by emphasising the rationale why SSG was used instead of natural speech or conventional speech synthesizers. In all of these studies, the cerebral processing of the auditory stimulus was measured using the mismatch negativity (MMN) component of the event-related potential and its magnetic counterpart (MMNm) elicited by deviant sounds occurring in a sequence of standard sounds (Näätänen, 1992).

In studies by Näätänen et al. (1997) and Cheour et al. (1998), both electrical and magnetic brain responses were compared between subjects of two closely related languages, Finnish and Estonian. It was shown that Finns generated smaller MMN and smaller MMNm left-hemispheric responses in comparison to Estonians to a stimulus that is a phoneme in Estonian but not in Finnish. An essential prerequisite feature of these studies was the total control of the formant structure of the stimuli according to the F1-F2-maps of the Finnish and Estonian vowels. In this study, SSG turned out to be a flexible method of stimulus generation due to its adaptive all-modelling of the vocal tract filtering that made generating vowels of a desired formant structure possible. The quality of the speech stimuli in Näätänen et al. (1997); Cheour et al. (1998) was close to natural and equal for both languages because the SSG-method used the identical glottal waveform as an excitation for all the synthesized vowels.

Shtyrov et al. (1998) analysed the effects of background noise on the cerebral functional asymmetry of speech perception. MMNm elicited by speech sounds presented in silence and during background white noise were measured. The main result was that speech stimuli registered in silence caused stronger mismatch responses in the left than in the right hemisphere, but when the stimuli was presented in noisy conditions, the activity of the left hemisphere diminished while that in the right hemisphere increased. Stimuli were two speech sounds that were acoustically different only during a very short period of time. Therefore, two plosive-vowel syllables ([pa] and [ka]) were produced with the SSG-method. The application of SSG was justified as follows. Firstly, SSG made it possible to use exactly the same glottal excitation for the vowel portion of the two syllables. Consequently the two sounds produced were acoustically equivalent during the vowel portion. This would not have been possible if natural speech signals were used. Secondly, it was very important to maintain the quality of the stimuli in silent conditions as close to natural as possible, because the authors were interested in analysing how hemispheric lateralization is affected when the sounds are distorted by additive noise. Adequate naturalness of short syllables would not have been achieved with conventional speech synthesizers.

3.2. Conclusions

This paper describes a novel, fully computerized method for the generation of speech stimuli in cognitive brain research. The method, semisynthetic speech generation (SSG), is based on the separated speech production model that simulates human speech production with 3 processes: the glottal excitation, the vocal tract filtering, and the lip-radiation effect. When synthesising speech with SSG, the first step is to estimate the glottal flow from a given speech sound using an inverse filtering method. The waveform obtained is then used as an excitation to a digital all-pole filter that artificially models the vocal tract filtering effect. The spectrum of the vocal tract model can be adaptively adjusted to generate a speech sound of a desired formant structure. The semi-synthetic sound is finally obtained by estimating the lip-radiation effect by differentiating the output of the vocal tract model.

SSG has been successfully used as a method of stimulus generation in several recent experiments. The most important benefit of the method in comparison with the commercial speech synthesizers is its ability to create speech signals that are not purely artificial due to the excitation process that is extracted from a real speech signal. To be able to transmit to the synthesized speech signal features that originate from the real human voice production mechanism improves the naturalness of the stimuli. The artificial processing is restricted only to the modelling of the vocal tract filtering effect. This allows one to exactly adjust the formant frequencies of the stimuli as desired, which is a feature that is typically needed in stimulus generation when auditory information processing of the brain is studied.

A synthesis method somewhat similar to SSG was used by Maiste et al. (1995). In their study, the categorical perception of two English syllables, [ba] and [da], was studied. There are two principal differences between SSG and the method used

by Maiste et al. (1995). Firstly, the formants of the stimuli in Maiste et al. (1995) were computed using a pitch-synchronous LPC-analysis based on the covariance method. This approach requires the exact position of the pitch period to be determined, which was done by using a special piece of equipment, the laryngograph (also called the electroglottograph). However, the laryngograph is typically not available in brain research laboratories, which makes implementation of this approach problematic. Secondly, the excitation signal used by Maiste et al. (1995) corresponded to the impulse-like residual signal given by LPC-analysis. It was stated that the resulting speech quality of the synthesized syllables was almost indistinguishable from the original sounds. This was probably the case because both of the syllables included the vowel [a]. However, if different vowels were synthesized using the same LPC-residual as an excitation, it is no longer possible to achieve natural quality because the LPC-residual depends on the vowel analysed and consequently, the same excitation waveform can not be used for high quality synthesis of different vowels.

SSG has been used so far in experiments where cerebral processing of phonemes or syllables has been analysed. However, the new method can be easily extended to be used to synthesize auditory stimuli of higher linguistic levels, for example, words or sentences. This comes from the fact that the adaptive computation of the glottal flow is performed in SSG using blocks of short duration (e.g. 20 ms). Therefore, in the case of a word, for example, that consists of several different phonemes the SSG-method is able to adjust to the acoustical changes of the sound. The all-pole modelling of the vocal tract in the synthesis stage of SSG can also be implemented in a block based manner if time trajectories of formants are required to be produced in synthesising sounds that comprise different phonemes. Since SSG estimates the true excitation of the vocal folds, the stimuli synthesized follow accurately the time trajectory of the fundamental frequency of the original speech, which implies that the prosodic features of speech are preserved when using SSG.

SSG takes advantage of computations that are widely used in modern digital engineering, like digital filtering and LPC-analysis. Therefore, the method can be constructed using commercial digital signal processing (DSP) software packages to meet the demands of, for example, cognitive brain research and maybe even certain clinical environments. The authors have implemented the method in the PC-environment using the QuickSig DSP-software (Karjalainen, 1990). This Lisp-language based implementation is available from the first author of this paper. In addition to the software, the user only needs a digitized speech sound as a source material for the computation of the glottal flow. Since inverse filtering is known to be sensitive for the quality of the recording equipment (Wong et al., 1979), digitising the speech sound should be done using a high quality condenser microphone and an audio card that has a flat amplitude response down to a few Hz.

Acknowledgements

This study was supported by the Academy of Finland.

References

Alku P. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Comm 1992;11(2-3):109–118.

Cheour M, Ceponiene R, Lehtokoski A, Luuk A, Allik J, Alho K, Näätänen R. Development of language-specific phoneme representations in the infant brain. Nature Neurosci 1998;1(5):351–353.

Fant G. The acoustic theory of speech production, The Hague: Mouton, 1960.

Flanagan J. Speech analysis, synthesis, and perception, New York: Springer, 1972.

Gold B, Rabiner LR. Analysis of digital and analog formant synthesizers. IEEE Trans Audio Electroacoust 1968;16(1):81–94.

Karjalainen M. DSP software integration by object-oriented programming: a case study of QuickSig. IEEE ASSP Mag 1990;7:21–31.

Kraus N, McGee T, Micco A, Sharma A, Carrell T, Nicol T. Mismatch negativity in school-age children to speech stimuli that are just perceptibly different. Electroenceph clin Neurophysiol 1993;88:123–130.

Kuriki S, Okita Y, Hirata Y. Source analysis of magnetic field responses from the human auditory cortex elicited by short speech sounds. Exp Brain Res 1995;104:144–152.

Maiste A, Wiens A, Hunt M, Scherg M, Picton T. Event-related potentials and the categorical perception of speech sounds. Ear Hear 1995;16(1):68–90.

Näätänen R, Lehtokoski A, Lennes M, Cheour M, Huotilainen M, Iivonen A, Vainio M, Alku P, Ilmoniemi RJ, Luuk A, Allik J, Sinkkonen J, Alho K. Language-specific phoneme representations revealed by electric and magnetic brain responses. Nature 1997;385:432–434.

Näätänen R. Attention and brain function, Hillsdale, New Jersey: Erlbaum, 1992.

Rabiner LR, Schafer RW. Digital processing of speech signals, New Jersey: Prentice-Hall, 1978.

Rogers R, Papanicolaou A, Baumann S, Saydjari C, Eisenberg H. Neuromagnetic evidence of a dynamic excitation pattern generating the N100 auditory response. Electroenceph clin Neurophysiol 1990; 77:237–240.

Sams M, Aulanko R, Aaltonen O, Näätänen R. Event-related potentials to infrequent changes in synthesized phonetic stimuli. J Cog Neurosci 1990;2(4):344–357.

Schröger E, Paavilainen P, Näätänen R. Mismatch negativity to changes in a continuous tone with regularly varying frequencies. Electroenceph clin Neurophysiol 1994;92:140–147.

Sharma A, Kraus N, McGee T, Nicol T. Developmental changes in P1 and N1 central auditory responses elicited by consonant-vowel syllables. Electroenceph clin Neurophysiol 1997;104:540–545.

Shtyrov Y, Kujala T, Ahveninen J, Tervaniemi M, Alku P, Ilmoniemi R, Näätänen R. Background acoustic noise and the hemispheric lateralization of speech processing in the human brain: magnetic mismatch negativity study. Neurosci Lett 1998;251:101–104.

Tiitinen H, May P, Reinikainen K, Näätänen R. Attentive novelty detection in humans is governed by pre-attentive sensory memory. Nature 1994;370:90–92.

Wong D, Markel J, Gray A. Least squares glottal inverse filtering from the acoustic speech waveform. IEEE Trans Acoustics Speech Sign Proc 1979;27(4):350–355.

Zatorre R, Evans A, Meyer E, Gjedde A. Lateralization of phonetic and pitch discrimination in speech processing. Science 1992;256:846–849.