Speech Communication 10 (1991) 497-502
North-Holland

Synthesis of breathy vowels: Some research methods

Dik J. Hermes

Institute for Perception Research/IPO, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Received 3 July 1991

Abstract. When vowels are synthesised by means of a source-filter model, a delta-pulse train is often used as a source signal. Although breathiness can to some extent be simulated by using a sophisticated glottal-source model, a more natural simulation of breathiness requires the addition of aspiration noise. When stationary noise is used, however, the noise is to a large extent perceived as coming from a separate sound source which hardly contributes to the breathy timbre of the vowel. This problem can be solved by using noise with a temporal envelope of the same periodicity as the pulse train. In a simple source-filter model, a combination of lowpass-filtered pulses and synchronous highpass-filtered noise bursts of equal energy was used as a source signal. In this way, the noise was no longer perceived as a separate sound, but integrated perceptually with the strictly periodic part of the signal. It will be shown that this integration consists of both a reduction of the loudness of the separate noise stream and a timbre change in the breathy vowel.

Zusammenfassung. Bei der Synthese von Vokalen mit einem Quellen-Filtermodell wird als Quellensignal oft eine Deltapulsfolge verwendet. Obwohl eine hauchige Stimme bis zu einem gewissen Grad mit einem verfeinerten Glottal-Quellenmodell simuliert werden kann, erfordert eine natürlich klingende Simulation doch die Addition von Rauschen. Wenn zu diesem Zweck allerdings stationäres Rauschen verwendet wird, wird dieses meist als von einer getrennten Signalquelle stammend wahrgenommen und trägt damit kaum zum Eindruck eines gehauchten Vokals bei. Dieses Problem kann dadurch gelöst werden, daß das Rauschen mit einer zeitlichen Einhüllenden versehen wird, deren Periodizität dieselbe ist wie die der Pulsfolge. In einem einfachen Quellen-Filtermodell wurde eine Kombination von tiefpaßgefilterten Pulsen und synchronen hochpaßgefilterten Rauschpulsen mit dergleichen Energie als Quellensignal verwendet. Dadurch wurde das Rauschen nicht mehr als getrenntes Signal wahrgenommen, sondern perzeptiv mit dem periodischen Signalanteil integriert. In dem Beitrag wird gezeigt, daß diese Verschmelzung einerseits auf einer Verringerung der Lautheit des Rauschsignals und andererseits auf einer Klangfarbenänderung im gehauchten Vokal beruht.

Résumé. La synthèse de voyelles au moyen d'un modèle source-filtre s'effectue souvent avec un train d'impulsions delta comme signal d'entrée. Bien que des modèles sophistiqués de la source glottale puissent être employés, dans une certaine mesure, afin de simuler une voix soufflée, une simulation plus naturelle exige l'addition d'un bruit d'aspiration. Cependant, lorsqu'un bruit stationnaire est employé, il sera perçu en bonne partie comme provenant d'une source sonore séparée qui ne contribue pas au timbre soufflé de la voyelle. Ce problème peut être résolu en utilisant un bruit ayant une enveloppe temporelle de la même périodicité que le train d'impulsions. Dans un simple modèle source-filtre, des impulsions filtrées passe-bas combinées à un bruit pulsatif synchrone filtré passe-haut à énergie égalisée ont été employées comme signal de source. De cette façon, le bruit n'est plus perçu séparément, mais est intégré perceptuellement à la partie strictement périodique du signal. Il sera démontré que cette intégration consiste à la fois en une réduction de la force sonore du bruit et en une altération du timbre de la voyelle soufflée.

Keywords. Speech synthesis, breathiness, vowel perception, auditory scene analysis.

1. Introduction

One of the problems one often encounters in the synthesis of complex sounds is that the components do not integrate perceptually, but stream into several sounds that seem to come from different sound sources. This problem expresses itself very clearly when one adds noise to a synthetic speech sound, e.g., for obtaining a more breathy sounding voice. If one uses stationary noise, this noise does not integrate perceptually with the speech sound, but continues to be perceived as a separate sound source and hardly affects the timbre of the speech signal.

0167-6393/91/$03.50 © 1991 Elsevier Science Publishers B.V. All rights reserved

These observations illustrate that signal components must fulfil certain specific conditions if they are to integrate with each other into one single auditory object having a timbre with characteristics derived from all components. In the past, a number of investigations have been carried out in this area (e.g. (Bregman, 1978; Darwin, 1981; Darwin and Sutherland, 1984; McAdams, 1984; Weintraub, 1985)), now referred to as "auditory scene analysis" (Bregman, 1990). A review of the implications of this field of research for speech perception is presented by Repp (1988).

Most of these investigations were concerned with tonal stimuli consisting of sums of pure sinewaves. Only Dannenbring and Bregman (1976) used combinations of noise and periodic sounds and conclude that the segregation effects of these stimuli were the strongest they had observed. Nevertheless, many natural sounds, including the human voice, do contain noisy components. Many attempts have been made to synthesise a natural-sounding breathy voice (e.g. (Makhoul et al., 1978), for a review see (Klatt and Klatt, 1990)). One of the problems, though not explicitly mentioned, is that the addition of too much noise results in the perceptual formation of a separate noise stream, which does not further increase the breathiness of the voice. In most cases, stationary noise was used to simulate aspiration noise, an exception being (Carlson et al., 1990).

In this paper, some conditions will be formulated under which noisy components do integrate perceptually with the strictly periodic part of the sound. The resulting sounds consist of breathy vowels, generated with a simple source-filter model. Depending on the extent to which the conditions for integration are fulfilled, a smaller or larger part of the noise does not integrate perceptually, but remains segregated from the vowel. Some experimental procedures will be proposed to investigate these findings more quantitatively. The implications of the results for speech perception and speech synthesis will be indicated.

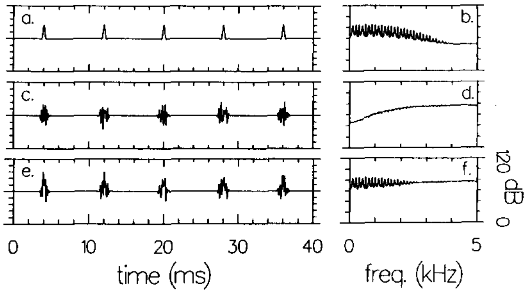

Fig. 1. The source signal for a cut-off frequency between periodic and noisy part of 2000 Hz. The waveform of the strictly periodic part is shown in (a), its long-term log-amplitude spectrum in (b). The waveform and the spectrum of the noisy part are shown in (c) and (d). Waveform and spectrum of the sum of (a) and (c) are shown in (e) and (f).

2. Synthesis

The composition of the source signal used for the synthesis of the breathy vowels is illustrated in Figure 1. The source is a combination of a strictly periodic part and a noisy part. The strictly periodic part is a lowpass-filtered pulse train. The waveform and the log-amplitude spectrum of such a pulse train are shown in Figure 1(a, b) for an example with a cut-off frequency of 2000 Hz. The period of the pulse train determines the pitch of the sound, 125 Hz in this case. The noisy part, shown in Figure 1(c, d), consists of a train of highpass-filtered noise bursts. The cut-off frequency is the same as that of the pulse train. So, pulses and noise bursts complement each other in the frequency domain. As a consequence, the complete source signal has a white spectral envelope. Waveform and log-amplitude spectrum of the source signal are shown in Figure 1(e, f). As each burst coincides in time with a pulse, the train of pulses runs in synchrony with the train of noise bursts. Since the noise consists of bursts, it is no longer stationary. Furthermore, the noise bursts have durations of less than one pitch period, *excessive peaks are removed* by applying a compressive nonlinearity, here an arctangent, and *all bursts have the same energy*. By passing this source signal through a combination of a *de-emphasis filter* and a *formant filter*, natural-sounding breathy vowels could be obtained. Figure 2 shows the waveforms and the log-amplitude spectra of

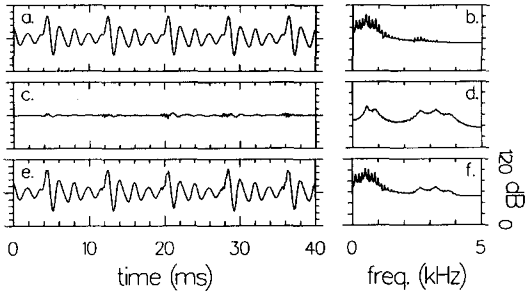

Fig. 2. Same signals as shown in Figure 1 after passing through the de-emphasis and the formant filters. The formants are those of the vowel /o/.

the strictly periodic part of such a vowel (Figure 2(a, b)), of its noisy part (Figure 2(c, d)), and of the vowel itself (Figure 2(e, f)).

The application of de-emphasis to the source signal was essential to obtain a natural-sounding breathy vowel. If no de-emphasis was applied, a large part of the noise segregated from the periodic part of the signal, producing a vowel with about the same timbre as when no noise was added. For a de-emphasis of 0.9, giving a spectral tilt of about 6dB/octave, however, the addition of the noise was clearly coupled with a timbre change in the direction of breathiness.

Breathiness appeared to depend much more on the cut-off frequency between the pulses and the noise than on the amount of added noise. For very low cut-off frequencies, below a few hundred Hz, so that the source signal consisted almost completely of noise bursts of equal energy, the synthesised vowels sounded so breathy that one was inclined to tell the computer to clear its throat. In that sense, these vowels still sounded natural.

3. Perceptual test

One of the problems in the study of perceptual integration and segregation is how to quantify the extent to which various components integrate. As far as the added noise is concerned, there are two sides to this problem. One side relates to segregation of the noise stream, the other to integration. In this study, it was assumed that the perceptual segregation of the noise stream can be determined by measuring its loudness. This amounts to assuming that the loudness of the noise stream decreases the more it integrates with the vowel. On the other hand, if noise integrates perceptually with the vowel, it will influence the timbre of that vowel. This means that, when the (high-frequency) noise contributes to the timbre of the vowel, i.e., when the noise is better integrated perceptually with the low-frequency periodic part, the timbre of the vowel will sound more like having a large high-frequency content. On the other hand, when the noise does not integrate, the timbre of the vowel will sound more like having a small high-frequency content.

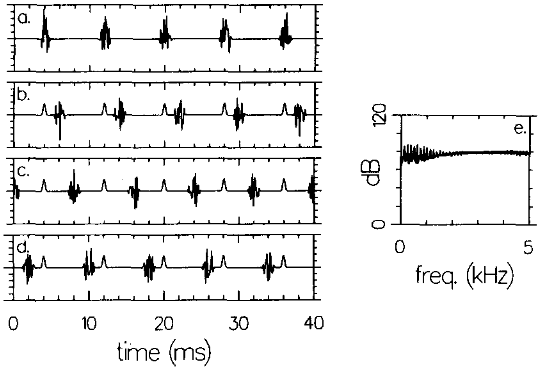

Fig. 3. Waveform and log-amplitude spectra of four of the eight source signals used for testing the integration and segregation of the high-frequency noisy part and the low-frequency periodic part. Observe that their long-term spectra are virtually identical.

The validity of these principles was tested in an experiment in which the importance of the relative temporal position of the noise bursts with respect to the periodic pulses was investigated. Figure 3 shows the waveforms and the log-amplitude spectra of four (of eight actually tested) source signals. (For reasons of clarity, these figures show stimuli synthesised with a sample frequency of 10kHz. In the actual experiments 20kHz stimuli were used.) Their waveforms are presented in Figure 3(a–d), while their long-term log-amplitude spectra are presented in Figure 3(e). The cut-off frequency between the noise and the pulses was 1200Hz, resulting in very breathy vowels. The pitch of the stimuli was 125Hz, which gives a pitch period of 8ms. Observe that the maximum distance between the pulses and the bursts is 4 ms. As can be seen in Figure 3(e), the log-amplitude spectra of these stimuli are virtually equal. In spite of this, it appeared that the high-frequency noise integrated better with the low-frequency periodic part of the stimulus, when the noise bursts coincided with the pulses. It was this aspect of the synthesis which was the object of the following two tests.

In order to measure the *segregation* of the noise, subjects were asked to adjust the variable intensity level of a comparison stimulus in such a way that its loudness was equal to the loudness of the noise stream in one of the test stimuli. The long-term log-amplitude spectra of the comparison stimuli with the lowest and the highest level are shown in Figure 4(a). In order to make their timbre resemble the timbre of the segregating noise stream, these comparison stimuli consisted of trains of high-pass filtered noise bursts of equal energy. This facilitated the task for the untrained subjects.

In order to measure the *integration* of the noise, subjects were asked to adjust the variable timbre of a comparison stimulus to the timbre of the vowel in one of the test stimuli. Figure 4(b) shows the long-term log-amplitude spectra of the two extreme source signals of the vowels used as comparison stimuli. These source signals consisted of trains of strictly periodic pulses, so that full perceptual integration into one vowel was guaranteed. As can be seen, these source signals differ in their high-frequency content. As mentioned, if the high-frequency noise integrates with the periodic part, the vowel will sound more like having a high-frequency content than when the noise does not integrate. Subjects found this experiment more difficult than the loudness matching. Therefore, all subjects were first asked to order the eight test stimuli according to timbre, and then to match them with a comparison stimulus. This made the task easier.

4. Results

The results of the two tests are presented in Figure 5. The abscissae give (in ms) the phase in the period of the pulse train at which the noise bursts were positioned. At zero phase, the maximum of the temporal envelope of the noise bursts coincides with the maximum of the pulse. The pitch period being 8 ms, the distance between the noise and the pulses is at most 4 ms. At phases higher than 4 ms, the bursts and the pulses come closer together, and run in synchrony, again, at a phase of 8 ms. So, the datapoints presented at 8 ms are the same as those presented at 0 ms. The ordinate in Figure 5(a) gives the average intensity level of the noise bursts in the comparison stimulus, as adjusted by 10 subjects. Vertical bars

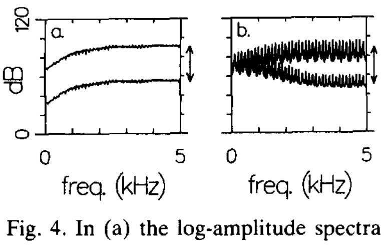

Fig. 4. In (a) the log-amplitude spectra are shown of two of the comparison stimuli, the variable level of which had to be adjusted in such a way that the loudness of this comparison stimulus matched the loudness of the noise stream in the synthetic vowel. The two spectra are those of the signals with the lowest and the highest intensities used. In (b) the log-amplitude spectra are shown of the source signals of two of the comparison stimuli, the variable high-frequency content of which had to be adjusted in such a way that the timbre of the vowel with this source signal best matched the timbre of the synthetic vowel. The two spectra are those of the source signals of the vowels with the lowest and the highest high-frequency content.

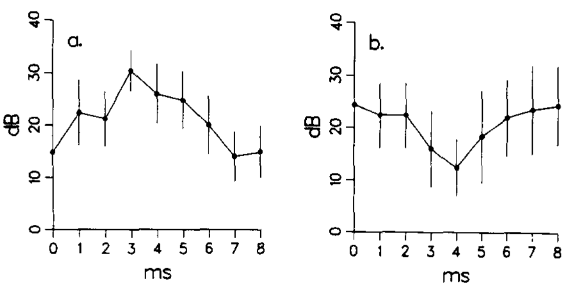

Fig. 5. The results of the two adjustment experiments. In (a) the average intensity of the comparison stimuli is shown for the experiment in which the variable intensity level of the noisy comparison stimulus was adjusted in such a way that its loudness matched the loudness of the segrating noise stream in the test stimulus. In (b) the average level of the high-frequency part of the pulse excited vowel used for comparison stimuli is shown for the experiment in which the timbre of the comparison stimulus was adjusted in such a way that its timbre best matched the timbre of the vowel used as test stimulus.

represent one standard deviation upwards and one standard deviation downwards. The ordinate in Figure 5(b) gives the average intensity level of the upper frequency region of the pulse-excited vowels, as adjusted by 10 subjects.

Figure 5(a) shows that the loudness of the noise stream is minimum when the interval between bursts and pulses is less than 1 ms. As the interval between bursts and pulses grows, the loudness of the noise stream grows, until it is matched to the loudness of a noise signal with an intensity which is over 15 dB higher than when the bursts and the pulses coincide. The maximum loudness of the noise stream is attained at an interval of 3 to 4 ms. After 4 ms, the interval between bursts and pulses decreases, again, which is coupled with a diminishing loudness of the noise stream. This shows that a much larger part of the noise does not integrate with the vowel when noise bursts and pulses do not coincide than when they do.

Figure 5(b) shows that, when noise bursts and pulses coincide, so at 0 and 8 ms, the timbre of the test stimulus has a relatively high-frequency content. As the interval between the pulses and the bursts increases, the timbre of the vowel looses something of its high-frequency content, until the adjusted level of the high-frequency content of the comparison stimulus is about 12 dB lower when bursts and pulses are maximally out of phase. This shows that, indeed, the noise stream as far as it integrates with the periodic stream contributes to the timbre of the vowel. For this experiment it has to be mentioned that the contributions of three subjects were not included. All three appeared to match the timbre of the test stimuli in exactly the opposite direction as the other subjects, when asked to order the test stimuli according to timbre. Two of these subjects were unable to proceed with matching the test stimuli to the comparison stimuli. Only the third completed these adjustments, and, as said, matched in exactly the opposite direction as the other subjects. The explanation for this is that he did not match the timbre of the vowel, but the timbre of the segregating noise stream with the comparison stimulus.

5. Discussion

The timbre of the comparison stimulus in the second test did not depend so much on the *absolute* level of the high-frequency part, as on the *relative* level of the high-frequency part with respect to the level of the low-frequency part. The ordinate in Figure 5(b) is, therefore, rather arbitrary, and only relates to the way in which the comparison stimuli were constructed.

The results shown here were obtained for a cut-off frequency of 1200 Hz. For these low cut-off frequencies, the vowel not only sounded breathy but also rather rough. This might indicate that across-critical-band phenomena such as comodulation masking release (Hall et al., 1984) and comodulation detection difference (McFadden, 1987) are responsible for these phenomena. These express themselves for modulation frequencies belonging to the range of roughness perception. Similar, though less strong, results could be obtained, however, for higher cut-off frequencies and pitches up to 300 Hz. This frequency more belongs to the range of pitch perception. The background of the phenomena described in this paper may, therefore, also be the same as that of the phenomena described by Duifhuis (1970, 1971), Bregman et al. (1985) and Wakefield and Viemeister (1985), which belong more to the field of pitch perception.

It might be argued that the decrease in loudness of the noise stream, when noise bursts and pulses coincide, might be due to partial masking of the noise by the pulses. While this may have played some role, it cannot explain the timbre change in the vowel. The incorporation of the noise into the timbre of the vowel demonstrates that the noise is not masked but can be heard actually, and integrates into the vowel stream.

The results show that noise can integrate perceptually with an otherwise strictly periodic signal, if it fulfils certain specific conditions. The noise should consist of bursts which run in synchrony with the periodic part of the sound signal. These noise bursts should not have excessive peaks and their energy in each pitch period should be about equal. If these conditions are not fulfilled, the noise does not integrate, or the vowels sound rough. From the speech production point of view, this might suggest that, at the phase in the glottal cycle at which noisy acoustic energy is produced, every resulting noise burst gets the same amount of acoustic energy. From the speech perception point of view, it appears that the temporal envelopes of the noise bursts within different frequency bands should have the same periodicity, while the amplitudes of the temporal envelopes should not fluctuate too much from period to period.

Acknowledgments

I would like to thank Jack Cullen and René Collier for their critical comments on the manuscript.

References

A.S. Bregman (1978), "The formation of auditory streams", in *Attention and Performance VII*, ed. by J. Requin (Erlbaum, Hillsdale, NJ), pp. 63–76.

A.S. Bregman (1990), *Auditory Scene Analysis* (MIT Press, Cambridge, MA).

A.S. Bregman, J. Abramson, P. Doehring and C.J. Darwin (1985), "Spectral integration based on common amplitude modulation", *Percept. Psychophys.*, Vol. 37, pp. 483–493.

R. Carlson, B. Granström and I. Karlsson (1990), "Experiments with voice modelling in speech synthesis", in *Proc. of the Tutorial and Research Workshop on Speaker Characterization in Speech Technology*, Edinburgh, 26–28 June 1990, ed. by CSTR (Edinburgh), pp. 28–39.

G.L. Dannenbring and A.S. Bregman (1976), "Stream segregation and the illusion of overlap", *J. Exp. Psychol.: Hum. Perc. Perform.*, Vol. 2, pp. 544–555.

C.J. Darwin (1981), "Perceptual grouping of speech components differing in fundamental frequency and onset-time", *Quarterly J. Exp. Psychol.*, Vol. 33A, pp. 185–207.

C.J. Darwin and N.S. Sutherland (1984), "Grouping frequency components of vowels: When is a harmonic not a harmonic", *Quarterly J. Exp. Psychol.*, Vol. 36A, pp. 193–208.

H. Duifhuis (1970), "Audibility of high harmonics in a periodic pulse", *J. Acoust. Soc. Amer.*, Vol. 48, pp. 888–893.

H. Duifhuis (1971), "Audibility of high harmonics in a periodic pulse. II. Time effects", *J. Acoust. Soc. Amer.*, Vol. 49, pp. 1155–1162.

J.W. Hall, M.P. Haggard and M.A. Fernandes (1984), "Detection in noise by spectrotemporal pattern analysis", *J. Acoust. Soc. Amer.*, Vol. 76, pp. 50–56.

D.H. Klatt and L.C. Klatt (1990), "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *J. Acoust. Soc. Amer.*, Vol. 87, pp. 820–857.

J. Makhoul, R. Vishwanathan, R. Schwartz and A.W.F. Higgins (1978), "A mixed-source model for speech compression and synthesis", *J. Acoust. Soc. Amer.*, Vol. 64, pp. 1577–1581.

S. McAdams (1984), Spectral fusion, spectral parsing and the formation of auditory images, Ph.D. Dissertation, Stanford University, USA.

D. McFadden (1987), "Comodulation detection differences using noise-band signals", *J. Acoust. Soc. Amer.*, Vol. 81, pp. 1519–1527.

B.H. Repp (1988), "Integration and segregation in speech perception", *Language and Speech*, Vol. 31, pp. 239–271.

G.H. Wakefield and N.F. Viemeister (1985), "Temporal interactions between pure tones and amplitude-modulated noise", *J. Acoust. Soc. Amer.*, Vol. 77, pp. 1535–1542.

M. Weintraub (1985), A theory and computational model of monaural auditory sound separation, Ph.D. Dissertation, Stanford University, USA.