Intelligibility and perceived quality of spectrally-impoverished speech sounds

PACS: 43.71.Gv

M. Paquier[1,2], G. Gilbert[2], J.C. Béra[1], C. Berger-Vachon[2], C. Micheyl[2,3]

[1] Laboratoire de Mécanique des Fluides et d'Acoustique CNRS UMR 5509
Ecole Centrale de Lyon, 36 av. Guy de Collongue, BP 163, 69131 Ecully Cedex, France

[2] Laboratoire "Neurosciences et Systèmes Sensoriels" CNRS UMR 5020
50 avenue Tony Garnier, 69366 Lyon Cedex 07, France
E-mail : mpaquier@olfac.univ-lyon1.fr, ggilbert@olfac.univ-lyon1.fr

[3] MRC – Cognition and Brain Sciences Unit
15 Chaucer Road, Cambridge CB2-2EF, United Kingdom

ABSTRACT

Ten listeners were asked to identify and rate the quality of speech sounds processed in such a way that only the most energetic spectral components were retained. The influence of both the number of retained spectral components and of the width of the frequency bands surrounding these components was studied. The sounds consisted of vowel-consonant-vowel stimuli produced by two male and two female speakers. Retaining five channels containing the most energetic spectral components was generally sufficient for the sounds to be correctly identified, and twenty frequency channels with a 20-Hz width proved sufficient to produce a "correct" sound quality.

INTRODUCTION

In certain situations, highly complex acoustic signals such as speech or music have to be transmitted through a limited number of spectral channels. This, for example, is the case with cochlear implants. Due to physical and physiological limitations, such implants deliver stimulus information to the auditory nervous structures via a limited number of electrodes, which convey signal components falling within different frequency bands. Current devices include up 22 electrodes, but often, the number of electrodes activated (quasi-) simultaneously is drastically reduced in order to avoid deleterious across-channel interferences, or simply because of technical limitations (e.g. energy consumption).

Under such circumstances, an important question regards the selection of the spectral components that are to transmitted by the device in order for the transmitted signal to retain as much relevant information as possible. A straightforward approach to this problem consists in choosing fixed stimulation channels to which fixed frequency bands are assigned. This approach has been adopted in certain cochlear-implant signal-processing strategies (like the CIS strategy of the Cochlear/Nucleus device). Although the choice of the electrodes and of the frequency bands assigned to them may be guided by knowledge on the general spectral properties of relevant signals, this approach is very likely to be sub-optimal since it does not

take into account the detailed spectral characteristics of the actual signals transmitted by the device. From this point of view, a better approach consists in selecting the spectral components that have the largest amplitude. Furthermore, since most environmental audio signals are characterized by a time-varying spectral content, it appears wise to update the selection of spectral components from one moment to the next.

This type of time-varying, amplitude-based selection of spectral components for transmission through limited-capacity devices forms the basis of signal-processing strategies implemented by current cochlear implants like the Nucleus 22/Cochlear corporation, with its SPEAK (spectral peak) strategy, or the Digisonic/MXM. The processor of these two devices divides the short-term spectrum of the signal into a fixed number of frequency bands, each of which is assigned to one available electrode on the electrode array. The amount of energy in each frequency band is then computed, and the devices offer the possibility that only the N electrodes corresponding to the N most energetic bands be stimulated. N is a parameter of the strategy, which can be pre-defined by the user. In the SPEAK strategy used with the Nucleus/Cochlear device, the value of this parameter is limited to 10, and is generally fixed to 6 or 8. In the Digisonic/MXM device, N can go up to 15 and assume two different values corresponding to two functioning modes of the device, which the user can select by means of a switch: a standard mode, in which N generally equals the number of functional electrodes, and a special mode, recommended in noisy environments, in which N is generally set to a lower value. Since stimulated electrodes are selected based on the short-term spectrum of the signal, corresponding to the temporal waveform segment comprised within the analysis window at the considered time point, the selected subset of electrodes stimulated can vary from one instant to the next.

Besides cochlear implants, the type of dynamic spectrally-based signal processing considered here might also in the near future inspire signal-coding strategies in digital acoustic hearing aids aimed at subjects with severe to profound hearing-loss. In these patients indeed, because frequency resolution is lost to a large extent, stimulation with only a limited number of widely-separated spectral components may be adequate. From this point of view, the selection of spectral components with the largest amplitude, at the exclusion of all other components, may be thought of as an extreme form of spectral-enhancement algorithms which aim to enlarge spectral contrasts (Boers, 1980; Summerfied *et al.*, 1985; Simpson *et al.*, 1990; Bunnel, 1990; Stone and Moore, 1992; Baer *et al.*, 1993).

An other question is the extent to which such dynamic spectral-reduction processing preserves the perceived quality of sounds. Such knowledge may be required in the near future since, with rapid technical and scientific advances in the field of auditory prostheses, the emphasis will probably shift from the mere preservation of speech intelligibility to the restitution of signals with the highest possible perceived quality. In other words, after allowing patients to understand speech to a reasonable extent through their auditory prostheses, researchers and designers will probably soon have to face the challenge of providing these patients with high-quality signals.

In this study, we examined the influence of two essential parameters of spectral-reduction signal-processing strategies on both the intelligibility and the perceived quality of speech sounds. The first parameter was the number of largest-amplitude spectral components retained. The second parameter was the width of the frequency bands corresponding to these components, which was directly related to the width of the FFT bins from which these components were extracted.

MATERIAL AND METHOD

Stimuli

Stimuli used in this experiment were VCV (vowel consonant vowel) signals. Choice of the vocalic context was done among a set of three vowels /a/, /i/ and /u/. The consonant was chosen among a set of 17 consonants of the French language. We obtained consequently 51 different stimuli pronounced 4 times by 4 different speaker (2 male, 2 female) which made a

corpus of 816 stimuli. The sequences were recorded in an anechoï c room, and were sampled at a rate of 44.1 kHz using a 16-bit dynamic range. These digital signals were further processed as follows: Firstly, the waveforms were divided into temporal segments of equal duration D, with consecutive segments overlapping over half of their duration. The signal contained in each segment was multiplied by a symmetric Hamming window. The resulting weighted signals were submitted to digital fast Fourier transform (FFT). For each segment, the N FFT components having the largest amplitude were selected; the amplitude of all the other components was set to zero. Finally, the modified spectra were submitted to an inverse FFT and the resulting temporal segments were added back together using the overlap-add technique (Allen, 1977) in order to reconstruct a temporal signal with the same duration as the original.

We were interested primarily in the influence of parameter N, the number of preserved FFT components, on the intelligibility and the quality of the reconstructed signals. N was varied between 1 and 160 (N=1, 3, 5, 10, 20, 40, 80, 160).

A second parameter of potential importance in which we were interested was the width of the FFT bins. When the overall analysis bandwidth is maintained constant, the total number of bins in the FFT is inversely related to the width of these bins. **t** is obvious that retaining 5 components in an FFT that contains only 10, each corresponding to a 100-Hz wide frequency band, is not the same as retaining 5 components in an FFT than contains 100, each 10-Hz wide. In the former case, 50% of the spectral information is preserved, in the latter, only 5% is. When the sampling frequency is kept constant, the width of FFT bins is controlled by the number of samples, or equivalently the duration D of the temporal segments. The four different temporal segment durations used in this study were 50 ms, 25 ms, 12.5 ms, 6.25 ms, and the respective corresponding FFT bin widths were 20 Hz, 40 Hz, 80 Hz, and 160 Hz. The use of different temporal segment durations presents the inconvenient that the segment duration determines the temporal rate at which spectral information is refreshed. Namely, when consecutive segments overlap over half of their duration, as was the case here, the updating rate of spectral information is equal to 2/D Hz (if D is expressed in seconds). With the signal-processing scheme described above, which involves the selection of FFT components with (potentially very) different frequencies each time the spectrum is refreshed, this may have a dramatic influence. In order to overcome this problem, we replaced each of several consecutive signal segments spanning a total duration of 25 ms (the longest single-segment duration used) with their average; this way, spectral information was effectively refreshed every 25 ms irrespective of segment duration.

Combining the eight N values indicated above with the four FFT-bin widths led to a total of 32 test conditions. These 32 conditions, multiplied by the 816 different recorded sequences, would lead to the presentation of 26112 successive stimuli per subject. Obviously, such a test was too long and was not feasible. In fact, the protocol consisted in presenting to subjects 2000 signals which are randomly chosen among the corpus. A preliminary study indicated that results was convergent as soon as we presented 1000 trials.

In the case where the width of FFT bins was equal to 160 Hz, having 160 bins would have led to an overall bandwidth of 25.6 kHz, larger than the Nyquist frequency (22.05 kHz). Therefore, in this condition, the actual number of spectral components retained was 136. This condition, in which the processing introduced no significant alteration in the original signal, was considered as a control, or reference condition.

Experimental protocol

Ten subjects (6 female and 4 male, 22 to 36 years old) with no known auditory problem took part in the study. Testing began with a short initial familiarization phase, the objective of which was to allow the listener to become acquainted with the experimental procedure, the different stimuli, and the range of damages undergone by these stimuli across the different processing conditions. This familiarization phase was followed by the actual test, which involved the successive presentation of the different processed sounds, in a pseudo-randomized order. Following the presentation of each stimulus, the subject had firstly to rate the perceived quality using one of the following six labels (as translated from French): “very degraded”, “degraded”, “correct”, “good”, “very good”. The following guidelines were provided: The “very good” label

was to be used only when the quality of the sounds being played was judged to be identical to that of a very good CD recording. The “good” label corresponded to the quality of a radio station with no musical character (for example news radios, which generally have very reduced bandwidths and strong compression ratios). The “degraded” and “correct” labels corresponded to the quality of a vinyl disc more or less damaged with a more or less ancient recording. Then, the subject had to identify the vowel and the consonant that had been presented. The test session lasted about two hours per subject. The digital signals were generated off line using a Pentium computer with Matlab v6.1. During the experiment, the signals were sent to a 16-bit digital-to-analog converter (EMU 10k1 DSP) of a Sound Blaster Live III card, using a sampling rate of 44.1 kHz. The stimuli were delivered diotically to the subject’s ears via Sennheiser HD265 Linear II headphones.

RESULTS

Figure 1 shows the quality scores, averaged across all subjects and all VCV, as a function of the number of FFT bins preserved and of bin width. The error bars represent the standard errors of the numerical quality scores averaged across subjects and runs.

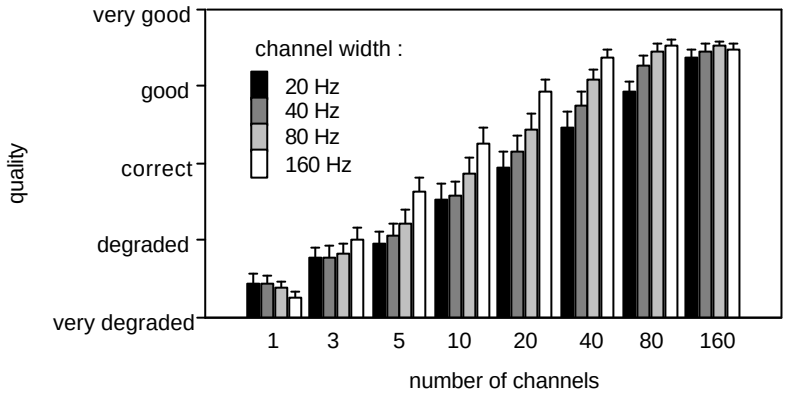

Figure 1: Quality judgments as a function of the number of retained spectral channels, with the channel width as a parameter.

These data were analysed using a two-way, repeated-measures ANOVA with the number of frequency components and the component width as factors. Both the number of retained components and the component width had a highly-significant influence on the estimated sound quality [$F(7,63)=645.185$, $p<0.0001$ for the former; $F(3,27)=169.125$, $p<0.0001$ for the latter]: the larger the number of retained components, and the wider these components, the higher the perceived quality score.

Figure 2 shows the percentage of errors, averaged across all VCV stimuli and all subjects, as a function of the number of retained FFT components and of the component width. The error bars represent the standard errors around the average quality scores (expressed numerically) across subjects and runs.

Incorrect identification of the consonant, the vowel, or both, was counted as one error. These data were analysed using a two-way, repeated-measures ANOVA with the number of spectral components and the component width as factors. Both factors had a highly-significant influence on the estimated sound quality [$F(7,63)=645.185$, $p<0.0001$ for the number of retained components; $F(3,27)=105.461$, $p<0.0001$ for the component width]: the larger the number of retained components, and the wider these components, the higher the intelligibility.

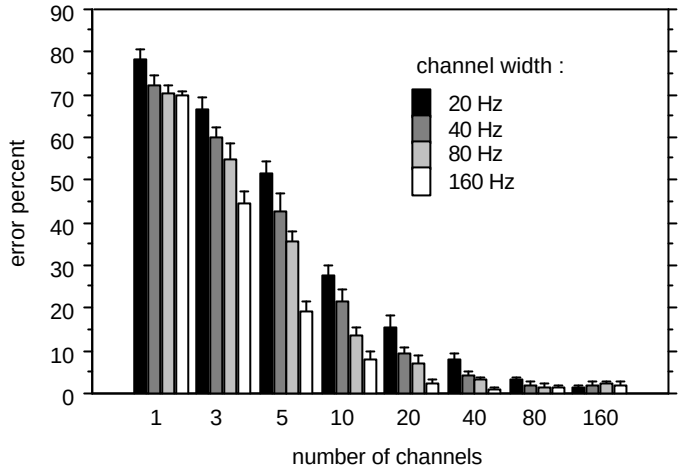

Figure 2: Error percentage as a function of the number of retained spectral channels, with the channel width as a parameter.

DISCUSSION

Quality

A striking result of the first task is that a relatively small number of largest-amplitude spectral components (N=20), each having a relatively narrow bandwidth (L=20 Hz), was sufficient for the quality of speech sounds to be judged on average as “correct”. This corresponds to an overall frequency range of only 400 Hz i.e. around 1/50th of the frequency range of normal hearing. Asymptotic quality scores were obtained with 80 components having each a 40-Hz width.

Overall, perceived quality scores were not related in a simple way to the overall bandwidth of the sounds: in general, higher quality scores were obtained using 2N channels having each a width L than using N channels of width 2L. This result may be explained by the fact that large amplitude components are predominant in determining sound quality and that large-amplitude spectral components are generally spread across the frequency range, rather than clustered together within narrow frequency ranges.

Increases in quality scores with increasing FFT-bin width were observed mainly when the number of retained bins was comprised between 5 and 40. When fewer components were retained, the quality rates were very low, whatever the bin width. When more components were retained, the quality was rated as "very good", whatever the bin width. This is consistent with the above-mentioned interpretation that the perceived quality of sounds depends primarily on how many of the large-amplitude spectral components are retained in the spectrum. When the number of channels was very small, increases in channel bandwidth were probably insufficient to ensure that a sufficient number of large amplitude components would be present in the spectrum of the processed sound. When the number of channels was very large, most maximum amplitude components were likely to be present in the spectrum of the processed sound already, so that further increases in bin width failed to produce a significant improvement.

Recognition

Retaining five components, each 20-Hz wide (which led to an overall spectral bandwidth of only 100 Hz) appeared to be sufficient for the VCV stimuli to be correctly identified in 50% of the cases. This finding agrees with results in literature which reveal that high speech recognition scores can be achieved with a small number of spectral channels, namely: around 46 for sentences and 8 for isolated words (Shannon et al., 1995; Dorman et al., 1997). The present result may also be paralleled with the finding that high levels of correct recognition can be achieved using sine-wave-speech, which is obtained by replacing the first three formants of an original speech signal by three sinusoids whose frequencies are equal to the formant centre frequencies (Remez et al., 1981).

Like the perceived quality score, the recognition score was not related in a simple way to the overall bandwidth of the sounds: retaining more components for signal reconstruction was generally more profitable than using wider components.

Increases in recognition scores with increases in the number of retained components and component width were mainly observed when the number of components was comprised between 3 and 40: when only one component was retained, recognition was obviously very difficult, and as soon as 80 or 160 channels were retained, the stimuli were always correctly identified.

Retaining forty frequency components with a 40-Hz width were sufficient to provide an error rate inferior to 5%. This corresponds to an overall frequency range of only 1600 Hz, i.e. less than 1/10th of the initial frequency range. This suggests that it may be possible to achieve a drastic reduction of frequency information whilst retaining very high levels of intelligibility.

CONCLUSION AND PERSPECTIVES

The present results suggest that the selection of narrow frequency bands around the spectral peaks is an appropriate approach for compressing audio signals, allowing to produce signals that are not only perfectly intelligible but also have a relatively high perceived quality. It may in particular prove more efficient than the use of a fixed, reduced bandwidth, which is still very common in current sound-processing devices (telephones, etc…). Provided that the transmitted spectral components are properly selected (on the basis of their large amplitude), both good quality and intelligibility can be achieved with a limited number of components, dispatched across the frequency range of hearing, and covering an overall bandwidth of only a few hundreds of Hz. These results may prove particularly useful in the optimization or design of signal-processing strategies for digital auditory prostheses, including in particular cochlear implants.

REFERENCES

Allen J.B. (1977). “Short-term spectral analysis, synthesis and modifications by discrete Fourier transform,” IEEE Trans. Acoust. Speech Sig. Proc. 25, 235-238.

Baer T., Moore B. C. J., Gatehouse S. (1993). “Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects of intelligibility, quality, and response times,” J. Rehab. Res. Devel., 30, 49-72.

Boers P. M. (1980). “Formant enhancement of speech for listeners with sensorineural hearing loss,” IPO Ann. Prog. Rep., 15, 21-28.

Bunnel, H. T. (1990). “On enhancement of spectral contrast in speech for hearing-impaired listeners,” J. Acoust. Soc. Am., 88, 2546-2556.

Dorman M. F., Loizou P. C., and Rainey D. (1997). “Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs,” J. Acoust. Soc. Am., 102, 2403-2411.

Green,D. M., and Swets, J. A. (1966). Signal detection theory and psychophysics (Wiley, New York).

Remez R. E., Rubin P. E., Pisoni D. B., and Carrell T. D. (1981). “Speech perception without traditional speech cues,” Science, 212, 947-949.

Shannon R. V., Zeng F. G., Kamath V., Wygonski J., and Ekelid M. (1995). “Speech recognition with primarily temporal cues,” Science, 270, 303-304.

Simpson, A. M., Moore, B. C. J., and Glasberg, B. R. (1990). "Spectral enhancement to improve the intelligibility of speech in noise for hearing-impaired listeners," Acta Otol. Suppl., 469, 101-107.

Stone, M. A., and Moore, B. C. J. (1992). “Spectral feature enhancement for people with sensorineural hearing impairment: effects on speech intelligibility and quality,” J. Rehab. Res. Devel., 29, 39-56.

Summerfield, A. Q., Foster, J., Tyler, R. S., and Bailey, P. J. (1985). “Influences of formant narrowing and auditory frequency selectivity on identification of place of articulation in stop consonants,” Speech Commun., 4, 213-229.

Warren, R. M., Riener, K. R., Bashford Jr, J. A., and Brubaker, B.S. (1995). “Spectral redundancy: intelligibility of sentences heard through narrow spectral slits,” Percept. Psychophys. 57, 175-182.