Envelope Information in Speech Processing: Acoustic-Phonetic Analysis vs. Auditory Figure-Ground Segregation.

Olivier Crouzet and William A. Ainsworth
Human and Machine Perception Research Centre
MacKay Institute of Communication and Neuroscience
School of Life Sciences, Keele University, Keele, ST5 5BG, United-Kingdom
{o.crouzet | w.a.ainsworth}@cns.keele.ac.uk

Abstract
Long-term envelope modulations (< 100Hz) influence the identification of speech in noise. It is not clear, however, whether this influence only takes place at the level of acoustic-phonetic analysis (phonetic identification) or if envelope fluctuations may also help in auditory figure-ground segregation (e.g. separation of speech from concurrent backgrounds). An experiment is presented in which the influence of long-term envelope modulations was investigated using signals mixed with either stationary or temporally modulated noise. The better performance observed when processing speech in modulated background may be related to the listeners’ ability to use envelope information in trying to follow concurrent signals independently. It is therefore predicted that, if long-term envelope modulations help to segregate speech from noisy backgrounds, this effect should be stronger when envelope information is fully available.

1. Introduction
Aperiodic complex sounds are characterized by various frequencies of the amplitude modulation spectrum, only some of which may be resolved as spectral components of the acoustic signals. Amplitude modulations within 100Hz and 20,000Hz are spectrally resolved by the human auditory system. They are characterised as spectral components of the acoustic signals which the human cochlea can analyse. Lower frequencies of the sound modulations (< 100Hz) fall below the spectral resolvability of the human auditory system. They reflect the temporal evolution of the amplitude envelope and, though they provide only global amplitude modulation, they have been shown to play a significant role in speech perception. Their effect has been studied using various speech signal manipulations consisting of either removing the long-term envelope modulations or suppressing the fine spectral structure while keeping these envelope modulations intact. Though the observed data are useful for an understanding of the effective cues for speech perception, it is not possible to reach a precise understanding of the various ways envelope information may be involved in processing speech.

2. Background
In this paper, it is argued that envelope information may be useful for both acoustic-phonetic analysis and auditory figure-ground segregation. Though previous data support the argument that amplitude envelope cues play a significant role in acoustic-phonetic analysis, they do not provide information with respect to the likely influence of envelope information on the segregation of speech from noise. The present experiment is aimed at investigating this issue.

2.1. Envelope information and the perception of speech in quiet
In order to study the influence of envelope information on speech perception processes, some authors have used spectrally degraded speech in which the amplitude of a broadband noise is modulated by the envelope of a reference speech signal [1, 2]. In these experiments, the role of envelope information has been studied with isolated utterances. Providing listeners with only envelope amplitude information does not completely hinder recognition. With VCV (Vowel-Consonant-Vowel) stimuli, Van Tasell et al. [1] showed that listeners are able to identify some of the phonetic features which define the medial consonant. Using everyday sentences, Shannon et al. [2] showed that normal-hearing listeners are able to reach 50% correct sentence recognition with a similar signal processing scheme with which only long-term amplitude modulations are available.

Though the data obtained with quiet speech lead to the argument that envelope information plays a significant role in acoustic-phonetic analysis, these modulations may also be useful for speech perception in noisy backgrounds. It is therefore straightforward to take a closer look at studies in which the influence of envelope modulations was investigated using noisy speech.
2.2. Envelope information and the perception of speech in noise

Using a different signal processing procedure, Drullman et al. [3, 4] also showed that envelope amplitude information is an important cue for the identification of speech. Rather than providing listeners with stimuli containing only temporal modulations with no fine spectral structure, they filtered the envelope of speech signals with various types of low-pass [3] and high-pass [4] filters while keeping the spectral structure intact. Though the resultant stimuli are easier to understand (and contain spectral information which may limit the influence of manipulated low-frequency modulations), this smearing enables one to investigate the effectiveness of various ranges of envelope modulation more precisely. To study the influence of modulation-spectrum filtering, stimuli were presented in a stationary background noise and the Signal-to-Noise Ratio (SNR) was measured using an adaptive procedure in which the background noise level was dynamically controlled. In these experiments, it is shown that Speech Recognition Thresholds (SRT) are lower when long-term envelope modulations are removed.

On the basis of these data, it is however not possible to dissociate the respective contribution of envelope information in acoustic-phonetic analysis and in the segregation of speech from noise. Indeed, filtering the envelope adds to the degradation that is produced by the addition of noise. As a matter of fact, the noise added in these studies was only a means to study the role of envelope amplitude modulation with a more precise control of their availability. However, it seems crucial to find a way of investigating the issue of their involvement in auditory figure-ground segregation.

2.3. Investigating the influence of envelope modulation on figure-ground segregation

The purpose of the present experiment is to investigate the specific involvement of long-term amplitude modulations on the recognition of speech signals in noisy backgrounds. By such, we mean not only their contribution to speech identification but their ability to provide information with respect to the process of matching each spectro-temporal component with its most likely acoustic source.

In normal-hearing listeners, the masking of speech by noise is highly dependent upon the temporal characteristics of the noise [5]. At an equivalent average signal-to-noise ratio, a stationary noise is more difficult to deal with than a time-varying noise. As envelope cues involve long-term modulation of the amplitude across time, they may account for this effect by providing listeners with cues to follow each acoustic source in the auditory scene. Therefore, if envelope cues are responsible for the facilitation of speech recognition in time-varying noise, smearing them should lower the performance of human listeners in speech-modulated noise but not (or more than) in stationary noise. Such a discrepancy between availability of envelope amplitude modulation and characteristics of the concurrent noise would provide a basis for arguing that they are useful for segregating speech from noise as specific cues to the auditory figure-ground segregation process and that they do not only provide cues to speech identification.

3. Method

In order to investigate the specific influence of long-term envelope modulation on the segregation of speech from noise, speech recognition performance was compared in stationary and modulated noise with two levels of high-pass envelope filtering of the target signals. It is predicted that stronger smearing of the envelope will hinder the advantage provided by modulated noise, therefore leading to a smaller effect of noise characteristics for greater high-pass cutoff frequencies.

3.1. Participants

Seven normal-hearing listeners, aged 20-30, took part in the experiment on a voluntary basis.

3.2. Stimuli

A subset of the TI-DIGIT database [6] was used. This database is made of continuous digit sequences pronounced by several American-English speakers in a quiet environment. The provided sound files are digitized at 20kHz (16 bit quantisation).

The selected stimuli were 4 or 5-digit utterances pronounced by both men and women. Filtering was performed off-line (see Section 3.3) within the MATLAB environment and the resulting stimuli were stored on disk. Two levels of high-pass envelope filtering were applied to the signals (0.5Hz, 64Hz).

Stimuli were presented by means of Sennheiser HD–435 headphones at a fixed average amplitude (approximately 65dB).

3.3. Signal processing

The signal processing scheme was similar to the one described in [4]. Stimuli were first passed through a 32-channel FIR filter-bank of approximately 1/At h octave bandwidth. The envelope modulation was then extracted within each frequency channel by means of a Hilbert transform and half-wave rectified to remove any negative values. The resulting envelope was under-sampled by a factor 64. High-pass filters (0.5Hz or 64Hz) were then applied to each frequency channel before up-sampling the envelope back to its original sampling frequency. The original signal was then multiplied by the ratio of the filtered to the original envelope. In order to prevent the oc-
Figure 1: Combination of two signals (black vs. gray) with full envelope modulation (left) and 64Hz high-pass envelope filtering (right). Stimuli processed through a 4-channel filter-bank.

occurrence of high-frequency aliasing, the resulting channels were low-pass filtered to the upper FIR filter-bank bandpass cutoffs before recombinig the 32 channels together. The final stimuli keep most of their fine spectral content but have lost the relevant long-term envelope amplitude modulations. Fig. 1 displays combinations of 4-channel amplitude envelope filtered signals passed through 0.5Hz (left) and 64Hz (right) high-pass filters.

The filtered signals were used to produce speech-modulated noise. Half the samples were negated on a random basis while the original value of the others was kept intact. The resulting stimuli can be described as a broadband noise modulated by the envelope of the reference signal. The long term spectrum of both modulated and stationary noises was dynamically adjusted to the concurrent signal’s spectrum during the experiment.

3.4. Procedure

The experiment took place in a sound-proof booth. Stimuli were transmitted through headphones and listeners had to provide each digit sequence they heard by means of a graphical interface displayed on a computer screen. A computerised PEST adaptive procedure (Parameter Estimation by Sequential Testing, [7]) controlled the outcome of the experiment by modulating the noise level with respect to the participants’ performance. For each condition, the session stopped when a satisfactory estimation of the threshold was reached. Each participant listened to the 4 signal–noise conditions (2 levels of envelope filtering, 2 types of noise). The order of presentation of both experimental conditions and target signals was randomized. At the beginning of each condition, the randomized lists were compared to prevent the combination of a target signal with a noise originating from this signal.

3.5. Results and discussion

A two–way ANOVA was conducted. The Signal-to-Noise ratio reached at the end of each adaptive session was set as the dependent variable. The data are displayed in Fig. 2. A strong effect of envelope filtering occurs. On the average, listeners’ threshold is at $-19.6\, dB\, SNR$ for the weak filtering condition but only reaches $-4.0\, dB\, SNR$ when most of the envelope modulations have been removed ($F(1, 6) = 78.1728, p < .01$). No global effect of noise characteristics is observed. Though a lower threshold is observed for modulated noise ($-12.3\, dB$) than for stationary noise ($-11.3\, dB$), this difference is only marginally significant ($F(1, 6) = 4.598, p = .076$).

However, the effect of noise modulations on listeners’ performance seem to vary with the amount of envelope filtering. At 0.5Hz cutoff, in line with the observations of Festen et al. [5], an advantage is observed for modulated noise. However, in the 64Hz high-pass cutoff condition, a better performance is observed for stationary noise. This interaction between envelope filtering
and noise characteristics reaches statistical significance ($F(1,6) = 7.762, p < .05$). In order to provide a better understanding of this effect, t-tests aimed at testing the effect of noise characteristics were applied on the data within each condition of the envelope filtering variable. This restricted analysis shows that the effect of noise modulations, though reversed (Fig. 2), is significant with both 64Hz ($t_b = 2.643, p < .05$) and 0.5Hz high-pass filtered signals ($t_b = 2.643, p < .05$). Therefore, while listeners processing speech signals with envelope information can benefit from the temporal modulation available in the concurrent background, removing parts of this information reverses the effect and makes perception easier in a stationary noise.

4. General Discussion

The experimental data confirm the hypothesis that long-term envelope modulations not only are useful for acoustic-phonetic matching but may also have a strong influence on auditory figure-ground segregation processes. By providing an alternation between acoustic components within each frequency band, they may help the speech processing system to follow acoustic sources more precisely due to the different long-term amplitude modulations which characterize each source. It is however not clear whether the observed effects may be interpreted in terms of cognitive processes aimed at using envelope information or if this effect may be accounted for by a greater acoustic masking of the spectral components when temporal envelope is smeared.

Whereas human listeners usually perform better in modulated backgrounds, Automatic Speech Recognition systems show a reversed pattern. This phenomenon must be related to the greater predictability of stationary noise.

As a matter of fact, noise reduction algorithms do not use envelope information yet. One may therefore compare our human participants listening to highly filtered speech with artificial systems processing acoustic signals with no processes designed to structure and predict the input on the basis of the envelope cues. This comparison seems useful to provide insights into the cognitive-acoustic issue. Indeed, when artificial systems process speech in modulated background, the envelope information is available in the acoustic input. Nevertheless, their pattern of performance do not replicate the human data. Therefore, spectral acoustic masking is not necessarily involved in reversing the stationary–modulated noise effect. It is crucial to investigate this issue more precisely to dissociate acoustic and cognitive explanations.

5. Acknowledgements

This work was supported by the European Community (SPHEAR, Speech Hearing and Recognition TMR Network). Special thanks are due E.F. Evans for his fruitfull advice about filterbank analysis.

6. References


