On the Various Influences of Envelope Information on the Perception of Speech in Adverse Conditions: An Analysis of Between-Channel Envelope Correlation.

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

Envelope information has been shown to influence speech identification in quiet. It is argued that long-term temporal modulation may also influence the processing of speech in both additive and convolutional noise. The ability of human listeners to use such information would be based on the correlation of long-term amplitude envelope information between spectral channels. The issue of whether such a correlation may be found in natural speech signals was investigated on a sample of 4-digit sequences extracted from the TI-DIGITS database. It is shown that the envelope of speech signals is highly correlated between spectral channels, especially when they are close to one another. The implications of this observation for the understanding of speech perception processes in adverse environments are discussed.

1. Introduction

It has been argued that monaural correlation between envelope amplitude in different spectral channels may play a role in auditory streaming by contributing to sound source determination [1]. Though the study of speech processing has provided data concerning the role of envelope *information* in phonemic identification, very little is known about the relationship between monaural envelope *correlation* and speech processing. The aim of this paper is to investigate the availability of envelope correlation in natural speech signals and to provide the basis to an investigation of its specific role for the processing of speech in adverse conditions.

1.1. Detection of across-channel synchrony

There is evidence that human listeners are able to perceive the correlation of amplitude modulated signals between spectral channels. When 2 to 5 sine-wave modulated narrow-band signals differing in centre frequency are produced with phase synchrony in one interval and without synchrony in the other one, listeners can discriminate between the intervals [2, 1]. This result provides evidence that the auditory system is able to process monaural envelope correlation, at least with a small number of sine-wave modulated signals in a simple discrimination task.

1.2. Envelope information and speech processing

Envelope information may influence the processing of speech signals in several ways. Though most of the work performed so far has focused on the investigation of the relationship between envelope amplitude and phonemic identification, influence on speech perception in multi-source environments or for processing desynchronized speech may also be predicted.

When fine temporal structure is removed without affecting the envelope amplitude modulation of a small number of frequency channels, some of the cues to phonemic identification remain. With VCV (Vowel-Consonant-Vowel) stimuli, listeners are able to identify some of the phonetic features which define the medial consonant [3]. A similar ability to identify spectrally degraded speech with only envelope amplitude information is found when using everyday sentences [4].

However, though the role of long-term envelope modulations for speech identification has been extensively investigated in the past few years, little is known about the role of across-channel correlation for processing speech with desynchronized spectral channels or in multi-source environments. If the availability of envelope correlation can help processing speech in adverse conditions, it is crucial to investigate the actual availability of envelope correlation in natural signals.

1.2.1. Extracting speech from noise

When concurrent acoustic signals are processed, their evolution in the spectro-temporal domain is independent. Each signal may however contain short-term spectral events with a common modification of amplitude across

Table 1: Centre frequency (and Bandwidth, in Hz) of each filterbank channel (the numbers on the left are the channel identifiers used in Tables 2 to 6).

1	132 (112)	***5***	1605 (814)
2	277 (177)	***6***	2590 (1155)
3	548 (366)	***7***	4102 (1871)
4	965 (467)	***8***	6427 (2779)

time. This comodulation may be helpful for stream formation. When the information processed by two auditory channels have a common long-term amplitude modulation, the auditory system may group these events together into a single auditory object whereas they would be parsed into different representations when the phase of their amplitude modulation is desynchronized.

1.2.2. Perceiving desynchronized speech

Adding noise to a signal is only one of several sources to the degradation of speech in natural environments. Convolutional noise like reverberation may also hinder recognition by changing the spectro-temporal structure of acoustic signals. However, human listeners can process speech in highly reverberant environments. Greenberg & Arai [5, 6] have shown that listeners are able to process sentences even when strong desynchronization is applied between spectral channels. Indeed, listeners can reach 75% correct recognition when processing sentences with up to 100 ms maximum delay between channels. Envelope amplitude correlation may also be useful to the auditory system in order to resynchronize the degraded temporal structure of fine spectral information.

2. Method

The aim of this analysis was to investigate the presence of a correlation of envelope information between spectral channels in natural speech signals. Due to the relationship between envelope amplitude and syllabic rhythm, it is expected that the correlation between spectral channels should be stronger for the lowest frequency envelope-modulation channels (0 - 4 Hz) than for higher ones (32 - 64 Hz). It is also expected that correlation coefficients should be stronger for close spectral bands, irrespectively of the low-frequency envelope modulations.

2.1. Material

Thirty 4-digit sequences uttered by male speakers were randomly selected from the TI-DIGITS database [7]. 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 20 kHz (16 bit quantisation).

2.2. Signal processing and analysis

Signal processing was performed within the MATLAB environment. Stimuli were first down-sampled by a factor of 64. They were then passed through an 8-channel FIR filterbank of approximately 1 octave bandwidth with few overlap between filters (cf. Table 1) Envelope amplitude was then extracted within each frequency channel by means of a Hilbert transform and half-wave rectified to remove any negative values.

Long-term envelope modulation channels were then selected by low-pass (0 - 4 Hz) or band-pass filtering (4 - 8 Hz, 8 - 16 Hz, 16 - 32 Hz, 32 - 64 Hz) of the resulting envelope.

Estimates of the correlation between the amplitude of envelope signals were performed by computing the Pearson product-moment correlation (r) between pairs of spectral bands within each modulation channel. Series of one-tailed t-tests were finally performed to check whether each of these coefficients were significantly greater than .50.

2.3. Results

Results are depicted in Tables 2 to 6. Correlation coefficients significantly greater than .50 are printed in slanted red. Non-significant coefficients are printed in light gray. Within the 0 - 4 Hz interval, a strong envelope correlation is observed between all channel pairs. Each of the computed coefficients is significantly higher than .50 ($p < .05$), which confirms the hypothesis that envelope amplitude correlation would be available in natural speech signals.

The observed correlation pattern differ for the remaining modulation channels (4 - 8 Hz to 32 - 64 Hz). Most of the observed coefficients do not reach the significance threshold. However, all coefficients along the diagonal prove to differ significantly from .50. This means that, though not all channel pairs exhibit a tendency to be comodulated in amplitude, pairs of adjacent channels are highly correlated to one another, whatever modulation channel is considered.

2.4. Discussion

As predicted at the beginning of this paper, a clear correlation of envelope information is observed between the spectral channels of natural speech signals. Though the correlation pattern differs between the 0 - 4 Hz interval and the other modulation channels, each of the analysed conditions shows a strong correlation between adjacent pairs of spectral bands. Moreover, focusing interest on the lowest part of the modulation spectrum which was investigated in this paper, a huge pattern of correlation is observed. It seems noteworthy to relate this observation to the vocalic amplitude rhythm of speech signals.

As clean natural speech signals contain phase infor-

Table 2: Mean envelope amplitude correlation observed between spectral bands. 0 - 4Hz modulation channel.

	2	3	4	5	6	7	8
1	***0.985***	***0.916***	***0.825***	***0.848***	***0.838***	***0.735***	***0.977***
2		***0.960***	***0.871***	***0.875***	***0.850***	***0.719***	***0.955***
3			***0.960***	***0.918***	***0.860***	***0.688***	***0.899***
4				***0.943***	***0.830***	***0.630***	***0.821***
5					***0.921***	***0.729***	***0.842***
6						***0.904***	***0.868***
7							***0.818***

Table 3: Mean envelope amplitude correlation. 4 - 8Hz modulation channels.

	2	3	4	5	6	7	8
1	***0.933***	***0.748***	***0.559***	0.512	0.529	0.339	***0.918***
2		***0.901***	***0.716***	***0.612***	***0.607***	0.374	***0.831***
3			***0.906***	***0.716***	***0.660***	0.364	***0.687***
4				***0.816***	***0.655***	0.323	0.537
5					***0.820***	0.396	0.497
6						***0.666***	0.566
7							0.450

Table 4: Mean envelope amplitude correlation. 8 - 16Hz modulation channel.

	2	3	4	5	6	7	8
1	***0.860***	***0.617***	0.481	0.425	0.388	0.296	***0.883***
2		***0.875***	***0.703***	0.553	0.506	0.390	***0.730***
3			***0.898***	***0.624***	0.559	0.401	***0.559***
4				***0.723***	0.553	0.355	0.486
5					***0.769***	0.465	0.442
6						***0.761***	0.491
7							0.419

Table 5: Mean envelope amplitude correlation. 16 - 32Hz modulation channel.

	2	3	4	5	6	7	8
1	***0.798***	0.509	0.394	0.310	0.332	0.304	***0.904***
2		***0.861***	***0.720***	0.476	0.462	0.404	***0.701***
3			***0.919***	0.542	0.494	0.406	0.469
4				***0.641***	0.514	0.379	0.393
5					***0.762***	0.496	0.302
6						***0.790***	0.411
7							0.423

Table 6: Mean envelope amplitude correlation. 32 - 64Hz modulation channel.

	2	3	4	5	6	7	8
1	***0.776***	0.479	0.360	0.248	0.283	0.249	***0.849***
2		***0.837***	***0.654***	0.410	0.397	0.338	***0.591***
3			***0.885***	0.508	0.423	0.332	0.365
4				***0.649***	0.467	0.318	0.315
5					***0.711***	0.403	0.216
6						***0.779***	0.397
7							0.466

mation with respect to the relationship between envelope amplitude modulation of more or less adjacent spectral bands, human listeners may prove to be able to use this information for processing degraded speech signals. This ability may be described both for additive and convolutional noise.

3. General Discussion

The results depicted in this study provide some insights into a better understanding of the influence of envelope information in speech recognition processes.

3.1. Speech processing in multi-source environments

As argued at the beginning of this paper (cf. Sec. 1.2.1), the availability of monaural envelope correlation may help to organize acoustic events into single auditory objects by providing a means to group events which were produced by a common source. This process may have been involved in the data presented in [8]. In this experiment, applying a 64 Hz high-pass filter to the envelope of speech signals prevented listeners from taking full advantage of the spectro-temporal dips available in mixtures of amplitude-modulated signals and favoured the identification of speech in stationary noise. With natural speech signals, stationary noise usually provides less information than modulated noise for the separation of speech from a concurrent background. Though this effect may also be caused by a smaller local Signal-to-Noise ratio when envelope modulations were smeared, it may be related to the absence of envelope correlation between the two concurrent acoustic sources which were mixed in the stimuli. With stationary noise, correlation was intact in the noisy part of the stimulus. An analysis of envelope correlation in envelope-filtered signals has to be performed.

3.2. Processing desynchronized speech

Involvement of monaural envelope correlation may also occur when processing desynchronized speech [5, 6]. Indeed, desynchronization induces a modification of the fine temporal-structure organisation between spectral channels. However, as natural speech signals exhibit a correlation of long-term temporal information across channels, listeners may use envelope information to resynchronize the fine spectral content between different channels. Envelope correlation would therefore provide a way to get access to the fast spectral modifications involved in consonant recognition. This ability may be based on various temporal-modulation channels but would certainly be stronger for the 0 - 4 Hz interval. As a preliminary observation to the comparison of our statistical analysis with human data, Fig. 1 depicts envelopes extracted from the two extreme channels investigated in this paper. In the left panel (0 - 4 Hz), envelope peaks

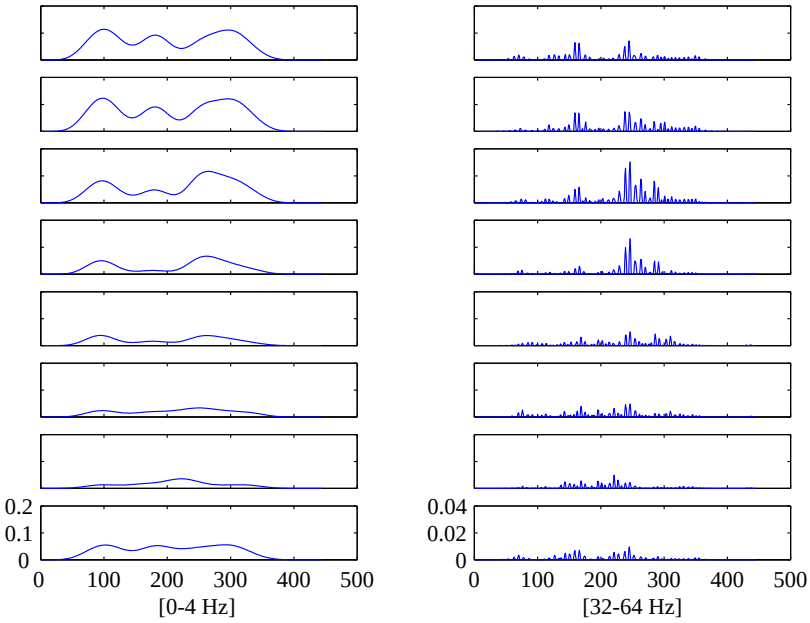

Figure 1: Filtered envelope amplitudes extracted from the utterance '9278' for 0–4 Hz (left) and 32–64 Hz (right) modulation channels (x-axis: time in ms, y-axis: envelope amplitude –due to large differences between modulation channels, units differ between the left and right parts of the graph–).

occur approximately every 100 ms. As a matter of fact, human performance starts degrading considerably above 100 ms delay [5]. When two peaks from alternate channels occur in phase, they may become more difficult to resynchronize.

3.3. Future work

Though the data presented in this paper does not prove that envelope information plays an effective role when processing speech in adverse conditions, they provide an important basis to the development of future work concerning the relationship between envelope correlation and speech recognition. As clean natural speech signals exhibit strong across-channel envelope correlation, it is now crucial to investigate the modification of these coefficients when speech is produced in additive or convolutional noise as well as with envelope filtered signals.

4. Acknowledgements

This work was supported by the European Community (SPHEAR, *SPeech HEAring and Recognition* TMR Network). Special thanks are due to G.F. Meyer for fruitful discussions about the underlying theory.

5. References

[1] J. Hall III and J. Grose, "Monaural envelope correlation perception in listeners with normal hearing and cochlear impairment," *Journal of Speech & Hearing Research*, vol. 36, 1993.

[2] V. Richards, "Monaural envelope correlation perception," *JASA*, vol. 82, pp. 1621–1630, 1987.

[3] D. Van Tasell, S. Soli, V. Kirby, and G. Widin, "Speech waveform envelope cues for consonant recognition," *JASA*, vol. 77, pp. 1069–1077, 1987.

[4] R. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, pp. 303–304, 1995.

[5] T. Arai and S. Greenberg, "Speech intelligibility in the presence of cross-channel spectral asynchrony," in *Proceedings of the ICASSP*, (Seattle, USA), 1998.

[6] S. Greenberg and T. Arai, "Speech intelligibility is highly tolerant of cross-channel spectral asynchrony," in *Proceedings of the Joint meeting of the ASA and the ICA*, pp. 2677–2678, 1998.

[7] R. Leonard, "A database for speaker-independent digit recognition," in *Proceedings of the ICASSP*, vol. 3, p. 42.11, 1984.

[8] O. Crouzet and W. Ainsworth, "Envelope information in speech processing: Acoustic-phonetic analysis vs. auditory figure-ground segregation," in *Proceedings of Eurospeech'2001*, (Aalborg, Denmark), 2001.