WAVELET TRANSFORM OF SPEECH AND STYLIZATION OF ACOUSTIC PARAMETERS

Z. Palková, J. Volín, J. Zimmermann

Institute of Phonetics, Faculty of Arts, Charles University in Prague
n. J. Palacha, 116 38 Prague, Czech Republic
zdena.palkova@ff.cuni.cz; jan.volin@ff.cuni.cz; julius.zimmermann@ff.cuni.cz

Abstract

One of the most crucial tasks in speech processing, whether in the psychological or the technological domain, is the division of the speech continuum into words. Many approaches have been tested to date, yet clear answers as to the foundations of the word recognition in speech are still missing. The present study employs a relatively new technological approach, the wavelet transform of the speech signal, to explore its potentials in helping to unravel the principles of word boundary identification. Based on the MATLAB wavelet toolkit, we designed an interface which provides clearly specified samples of spectral bands of the speech signal for further perceptual testing. The wavelet based decomposition and resynthesis seem to provide encouragingly unrestrained insight into the composition of the speech signal. The importance of individual spectral sections has been tested and confronted with the temporal structure of energy patterning to hypothesize about individual listeners' perceptual strategies.

1 Introduction

Current speech science, as defined e.g., in [1, 2], comprises three major components: speech engineering, psychology and linguistics. Although the experts from the three fields still find it sometimes difficult to talk to each other in a meaningful manner, we are learning to discuss problems more effectively than in the past. One of the reasons is the occurrence of interesting findings at the border area of the three disciplines; the other is the existence of the common objective: modelling correspondences between continuous acoustic events and discrete meanings in language, even if from different perspectives. Researchers of the past probably did not realize that the task of modelling speech behaviour would turn out to be so complex. The multifaceted interactions between several structural layers in frequency and amplitude domains (including interactions between these domains) together with contextual dependence of phenomena in temporal chains make the issue knotty, yet it is the role of the preceding knowledge (experience) of the language user that complicates the matter beyond our current explanatory capacity. The additional complexity is added by potential affective states of the speakers, which reflect their moods, attitudes, emotions, interpersonal stances or more permanent affective features of their personalities. To ignore affective states would mean to miss some very interesting applicational possibilities (e.g., [3, 4]).

However, for the present task, we have to abstract from speaking styles and affective states. Yet even for neutral, unmarked utterances, it is reasonable to assume that the core unit of the linguistic code is the word. This assumption is supported by quite a substantial body of research, but even more practically by the fact that the accumulated knowledge of humans lead in most orthographies to the same result: words are captured as single units separated from each other in a distinct manner. In contrast, other units of the linguistic structure may either not be graphically separated or if they are, then not in an unambiguous fashion. The existing algorithms of automatic speech recognition (ASR) assemble words from a complex of parameters which most probably differ from those used by human perceptual mechanisms. Rather than building perfect automatic recognizers, we see our task in exploring the nature of human perception of speech. The underlying method of our inquiry then relies on isolating and reassembling individual perceptual cues to establish their role in word recognition. The main objective of the current experiment it to particularize the method in some procedural detail.

Our previous research indicated that apart from semantic (and syntactic) context, words are recognized in the speech continuum with the help of specific F0 contours [5] and/or temporal configurations [6]. This view resonates with a large body of international research. Therefore, we decided to extract limited spectral bands of natural speech with specific F0 and temporal features, but without semantics (i.e., without information about the segmental make-up of the words) and carry out perceptual testing of the impact of the selected extracts on the listeners. The excision of the spectral band was performed through the wavelet transform of speech. This choice deserves a few notes concerning advantages of the wavelet transform over the more traditional techniques of signal processing.

First, the wavelet transform (WT) of speech can be considered a multiresolution analysis, in that it does not compromise between time and frequency domains (like the Fourier transform) and has better frequency resolution for lower frequencies in the effective range of speech, while the higher bands are analyzed with better time resolution [7, 8]. Second, unlike classical filtering, the wavelet decomposition of the speech signal does not double bits of information, hence it is more economical. The number of the output coefficients in WT is the same as the number of the input samples (see Fig. 1).

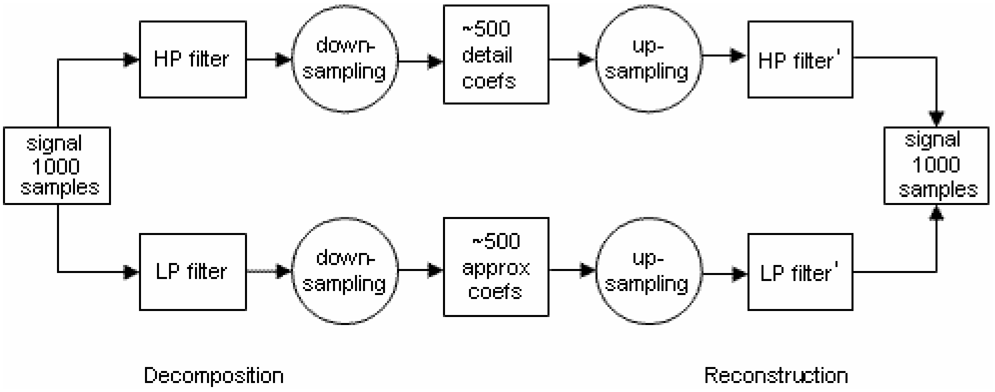


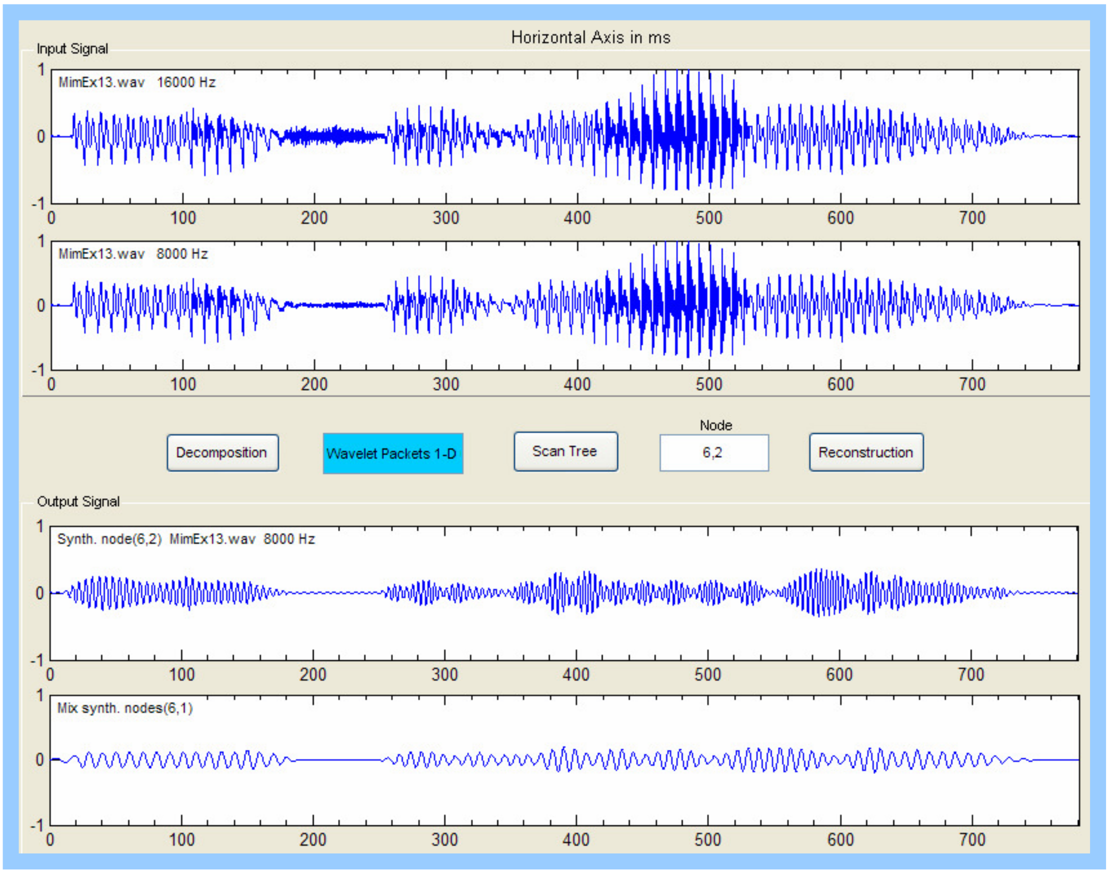
Figure 1. Diagram of the wavelet decomposition and reconstruction of the speech signal specifying the informational economy.

Third, the utilization of the mirror quadratic filters in the process guarantees the perfect reconstruction of the signal at any node of the analysis. It follows that any component of the decomposed signal can be removed or modified and provide an opportunity to test their role in the signal. Similarly, individual components can be convoluted without any artefacts. Such capacity offers interesting experimental conditions, which we decided to exploit. Our task was set to utilize recordings of ambiguous syllable chains and test the perception of the components that were identified as carriers of prosodic information.

2 Method

The instrumental platform of our experiments relied on MATLAB Wavelet Toolbox with a wide choice of programmable parameters, including a selection of mother-wavelet shapes, which, up to date, were identified as useful for speech processing. However, the GUI-1D of

MATLAB that is quite logically suited for speech signals in principle, was lacking certain essential features and had to be reprogrammed to allow replaying, saving and reconstructing the samples of speech. Figure 2 provides an example of our graphic interface. The top part presents the waveform of the signal, while immediately below is the resampled version (sampling frequency can be set to any meaningful value). Both waveforms can be replayed and compared auditorily. The waveforms in the bottom part of the screen show selected components that were extracted to be tested. These, again, can be replayed, saved, but in addition, they can be convoluted. The buttons in the middle of the screen start the decomposition of the signal, scanning the analysis tree so that individual nodes can be selected and explored and reconstruction of the signal from the components that were chosen. Experimenting with our speech material led us to believe that the sixth level of analysis provided perceptually most promising bands of speech signal.



Node
6,2

Figure 2. An example of the interface created to control the wavelet decomposition and reconstruction of the speech signal in the MATLAB environment.

The speech material consisted of oppositions of four-syllable sequences comprising the same segments but two different meanings dependent on perceptual solutions. Each four-syllable sequence could be interpreted as one word, or two or more words. For instance, the sequence [n e s e z n a: m i: m] could be interpreted as *nese známým* (*is carrying to acquaintances*) or *neseznámím* (*I will not get acquainted*). Such sequences were built into meaningful paragraphs which were read by three native speakers of Czech. The speakers were asked to read the carrier paragraphs in natural, fluent manner. They were not aware of the target items and did not know the purpose of the recording.

The individual items were consequently taken out of their contexts and were presented to groups of listeners to see, whether they can recognize the intention of the speakers (one-word or more-word option) without the context. Only the items with reasonable agreement among the listeners on the meaning were further processed.

Each selected item was extracted and transformed to retrieve the nodes 6,3 (third band on the sixth level of analysis) and 6,6. I addition, the aggregate of the two nodes was created. The perceptual test consisted of 18 items, ten of which were repeated. The resulting 28 items were balanced as to the linguistic properties of the material: one half contained whole word solutions, the other contained solutions consisting of more than one word.

The participants of the testing received an answer sheet with graphical representations of the stylized contours. The two-dimensional representations captured the movements of F0 and durational structure of the item. Each item was represented by four contours, one of which was a true representation. The others were: the melodic reversal of the true contour, the temporal reversal of the true contour, and the representation of the contour competing linguistically with the true one.

The stylization of the contours was carried out in the following manner. Individual syllabic peaks were manually labeled and the minimal durations and minimal mean F0 values were identified. Next, the relationship of the remaining three syllable peaks was established using the algorithm (1) for F0 and (2) for durations:

$$y_i = \text{if}(x<(\text{min}*1.05);\text{"L"};\text{if}(x<(\text{min}*1.1);\text{"M"};\text{"H"})) \qquad (1)$$

$$y_i = \text{if}(\text{y}<(\text{min}*1.5);\text{"Short"};\text{"Long"}) \qquad (2)$$

It is obvious, the F0 stylization utilizes 3 levels: high (H), mid (M) and low (L), while durational structures is captured in two levels only: short and long.

All the items were normalized for the listeners to -3dB level and were played through quality headphones in the quiet room. Each item was played five times followed by 15-second stretch comprising silence, desensitization, silence again, and a beep to alert the listener before the next item. The listeners watched the four options on their answer sheets and after hearing the item, they circled the contour they believed was the true one. Altogether, we collected 196 judgments: 7 listeners × 28 test items.

3 Results

Given that the multiple-choice format offered four options, the chance level was 25 %. Our respondents provided 60.2 % of correct answers. This result is high above the chance level and confirms that the method we used is valid, or at least promising in that it reflects perceptual capacities of human listeners.

Apart from the overall success rate, we were interested in the nature of confusions and the strengths of agreements among the respondents. This situation is captured in Table 1, where the results are expressed in the numbers of judgments in favor of the individual test options and with regard to the number of listeners agreeing with a particular choice for the given item.

It is obvious that agreements of five or more listeners occurred only in the case of correct options (the ‘true’ column). Temporal reversal (t-Rev. column), on the other hand, was the least favoured option. Only about 9 percent of the judgments were allocated there. F0 reversals attracted the listeners in a slightly greater number of occasions and the most frequent incorrect answer was that of a linguistically competing item.

n of 7	True	Comp.	F0-Rev.	t-Rev.
chosen 0	0	0	0	0
chosen 1	0	9	5	10
agreement 2	8	12	10	8
agreement 3	21	6	6	0
agreement 4	16	8	4	0
agreement 5	35	0	0	0
agreement 6	24	0	0	0
agreement 7	14	0	0	0

Table 1. The levels of agreement amongst seven respondents on individual items expressed in numbers of judgments awarded.

Table 2 displays the success rates for individual acoustic conditions. The node 6,3 captures the spectral band roughly between 170 and 260 Hz (adjusted according to the mean F0 of the individual speakers), whereas the node 6,6 implies the band between about 360 and 440 Hz. The line 6,3+6,6 in the table refers to the aggregate of the two bands. It can be seen that the lowest band was the most successful while the greatest number of errors occurred in the aggregate of the two spectral bands. The band of the harmonics in the first formant field provided mediocre results. In sum, the area of F0 and the first harmonic provided more information than the other two conditions. The success rate of 71 percent massively exceeds the chance level.

node	*n* items	correct	% corr.
6,3	70	50	71
6,6	63	37	59
6,3+6,6	63	31	49

Table 2. The success rates for individual analysis nodes used in the perceptual testing.

Although we had no expectations about the role of the linguistic content in the perception of stylized contours, we still hoped that some pattern might emerge during the testing. This did not happen. The success rate for whole-word items was about 58 percent while the fragmented items were evaluated correctly in about 62 percent of the cases. Obviously, the difference was verified as statistically insignificant (χ^2 (1; n = 196) = 0.34; p = 0,56).

4 Conclusion

The results of the experiment confirmed that the method of contour stylization based on the wavelet transform of speech leads to interpretable results and can, therefore, be considered, a valid method of experimenting. The programmable MATLAB environment provides a suitable platform for test item preparation, mainly by allowing for decomposition and perfect reconstruction of the speech signal.

The fact that the greatest number of errors occurred in the aggregate of the two spectral bands may be explained by realizing that although the signal contained more information than its individual components, the information was sometimes conflicting. Given the lack of other spectral components which could have served to establish coherent acoustic scene, the listeners had to guess rather than decide.

The area of F0 and the first harmonic proved to be more informative the other two conditions. This band is clearly the source of our perceptual cues with regard to prosodic structuring. However, it has to be admitted that the test task was highly abstract and did not allow us to make any conclusions concerning the relationship between the linguistic content and the perceptual effects. This is where the future work should be focused. We can still conclude that the most common confusions in the perception task between the “whole-word” and “more-word” solutions suggest the importance the fine phonetic detail, which was destroyed by the processing of the testing items [see, e.g., 9]. While the recordings with the preserved phonetic detail were in general not confused by the listeners to a great degree, the absence of the detail led to erroneous perception [10].

Acknowledgements:

This research was supported by the European Union grant MRTN-CT-2006-035561 – *Sound to Sense*, and by the Czech Ministry of Education grant VZ MSM0021620825.

References

[1] Skarnitzl R., Machač P.: Kohlerovy úvodníky a vize české fonetiky. Naše řeč 92, pp. 263–267, 2009.

[2] Volín, J.: Stability and dynamism of speech science terminology. In: I. Dominiková a M. Lachout (Eds.), Lingua Terminologica, pp. 106–115. MUP, Praha 2010.

[3] Bořil, H., Sangwan, A., Hasan, T., Hansen, J. H. L.: Automatic excitement-level detection for sports highlights generation. In: Proceedings of Interspeech'10, 2202–2205. ISCA,Makuhari, 2010.

[4] Bořil, H., Sadjadi, O., Kleinschmidt, T., Hansen, J. H. L.: Analysis and detection of cognitive load and frustration in drivers’ speech. In: Proceedings of Interspeech'10, 502–505. ISCA,Makuhari, 2010.

[5] Palková, Z., Volín, J.: The role of F0 contours in determining foot boundaries in Czech. In: Proceedings of the 15th International Congress of Phonetic Sciences, pp. 1783–1786. IPA&UAB, Barcelona 2003.

[6] Volín, J.: On the significance of the temporal structuring of speech. In: M. Malá, P. Šaldová (Eds.) ...for thy speech bewrayeth thee, pp. 289–305. UK FF, Praha 2010.

[7] Zimmermann, J.: Spektrografická a škálografická analýza akustického rečového signálu. Náuka, Prešov 2002.

[8] Torrence, C., Compo, G.P.: Practical Guide to Wavelet Analysis. Bulletin of the American Meteorological Society 79/1, pp. 61–78, 1998.

[9] Hawkins, S.: Roles and representations of systematic fine phonetic detail in speech understanding. Journal of Phonetics 31, pp. 373–405, 2003.

[10] Palková, Z.: The influence of context on perception of the prosodic word. In: R. Skarnitzl (Ed.) Phonetica Pragensia XXII (AUC-Philologica 2009/1), pp. 7–20, Karolinum, Praha 2010.