RESULTS ON PERCEPTUAL INVARIANTS TO TRANSFORMATIONS ON SPEECH.

Arnaud Robert

Electrical Department, Swiss Federal Institute of Technology, Lausanne, Switzerland.
e_mail: robert@circ.epfl.ch

ABSTRACT

This paper presents results of a study on perceptual invariants to transformations on the speech signal. A set of psychoacoustic tests were conducted as to put forward these invariants for the human hearing system (HS). The starting point is the decomposition of speech by an AM-FM analysis, rather than the use of more standard analysis methods. The main result of this work is the finding that our HS is robust to - namely our perception is not altered by - instantaneous frequency (IF) changes within a certain range, even though these resulted in substantial waveform modifications. This stimulated us to conduct further study on how standard analysis methods would cope with perceptually invariant changes; results show that, in fact, they are not robust to such changes. Finally, some applications of IF changes are proposed.

1. INTRODUCTION

Most speech applications are based on *linear* production models and on short-term spectrum analysis of the signal, usually extracting the envelope of the short-term spectrum. These include linear prediction and cepstral analysis (see [3] for a good review).

On the other hand, any wide-band signal can be fully represented by the envelopes and phases of narrow-band decomposition of the signal, given that the filtering process covers ideally the full frequency range of the signal. This is done in our HS at cochlear level where incoming sounds are broken up into many narrow-band signals each transmitted in separate narrow-band channels. This time-varying instantaneous frequency and envelope (or AM-FM) representation can describe nonlinear and time-varying phenomena occurring at speech production level, in agreement with experimental evidence in support of *nonlinear* model of speech production [5, 4]. As an example, the AM-FM representation was successfully used by Maragos in the fields of speech synthesis and coding [2].

A question that occurred to us is how precise must this phase-envelope representation be for our HS, or in other words to what extent do some modifications on this information lead to changes in perception. To provide an answer, we conducted psychoacoustic tests evaluating ranges of perceptual invariants.

The paper is divided as follows. In section 2 we describe the different stages of our experiments - including processing and transformations on the signal. Results of our psychoacoustic tests are presented in section 3. Finally, in section 4, a discussion on the results is addressed and conclusions are drawn.

The author would like to thank Dr Dedieu for his helpful discussions and the listeners of the psychoacoustic tests. This work is supported by the Swiss National Scientific Foundation, grant number 2150-045689.95.

2. METHODS

The complete processing chain is given at figure 1. First, the input signal $s(t)$ is decomposed by a band-pass filter bank. Each sub-signal is transformed into an AM-FM representation by H and then modified by a function T. All are recombined to give a new version of the original signal, $s'(t)$. A perceptual invariant is found if our HS is unable to differentiate the two signals. Let us now detail the different stages.

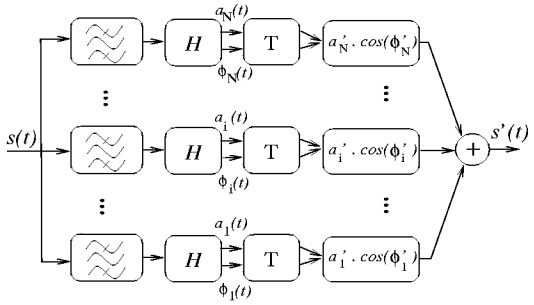

Figure 1: Overview of the processing method.

2.1. Processing

The first processing stage is an auditory-based filter bank, composed of N=16 filters with centre frequencies (CF) in the range 100-3400 Hz. The CF and bandwidth distributions were set according to the BARK scale, given in [3]. FIR filters of order 500 were used.

The second stage computes the Hilbert transform of each sub-band signal, providing an AM-FM representation. The Hilbert transform has been preferred to other possible decomposition schemes [4] because of its easy implementation. At this level, the speech signal can be written as the

sum of narrow-band time-varying scaled cosines:

$$s(t) = \sum_{n=1}^{N} s_n(t) = \sum_{n} a_n(t) \cdot cos(\ \phi_n(t)\ ) \qquad (1)$$

In the next step, we modified the envelope and phase of each narrow-band signal s_n by both scaling and temporal shifts operations. The used schemes are presented in table 1; parameters $\delta ta, \alpha, \delta t\phi$ and β are constant if not otherwise mentioned. Hence, at the output of the system, we have:

$$s'(t) = \sum_{n} a'_n(t) \cdot cos(\ \phi'_n(t)\ ) \qquad (2)$$

T1	T2	T3	T4
$a_n(t + \delta ta_n)$	$\alpha_n \cdot a_n$	$\phi(t + \delta t\phi_n)$	$\beta_n \cdot \phi_n$

Table 1: Transformation schemes.

2.2. Psychoacoustic tests

In order to determine if the transformations described above were perceptible (or not) by our HS, we conducted psychoacoustic tests involving *ten* listeners. The testing conditions were the following: listeners changed the parameter being tested (α, β,...) until they heard a degradation between original and modified speech, both of which could be played individually or successively at all times. Listeners were told that degradation meant either the presence of an audible noise or a loss of intelligibility, which ever bothered them first. To ensure better concentration of the listeners, tests were divided into envelope and phase modifications and each lasted 10-15 minutes at the most. Each test was composed of 4 to 5 utterances, single words or combination of words. Samples were played through high quality headphones in a low-noise room.

Although this testing method may be criticised, we believe it is adequate to show that some parameters can be changed by large amount without provoking any degradation to the signal's perception.

2.3. Implementation

Hilbert transform and parameters changes were implemented using MATLAB^{TM} functions and C programming. The speech segments were taken from the telephone quality PHONEBOOK database, with sampling rate of 8000 Hz and coded on 16 bits.

3. RESULTS

In this section, we will present the results of the four testing schemes T1-T4, with an emphasis on the one that revealed the most interesting results, T4. The ranges over which parameter changes were not perceptible by our listeners are reported in table 2. For all figures presented below, the utterance is the word *"accumulation"* spoken by an English male.

Before anything, we verified that the processing chain without any parameter changes lead to perceptually equivalent sound signals, which listeners all agreed on.

T1	$0 < \delta ta < 4ms$	T2	$0 < \alpha < 1 \rightarrow X$
T3	$0 < \delta t\phi < 2\pi$	T4	$-0.35 < \beta < 0.6$

Table 2: Psychoacoustic tests results.

3.1. Results of tests T1-T4

Envelope: tests T1 & T2 . Test T1 revealed that the envelope timing is essential within a certain range, which was expected since timing plays an important role in HS processing [1]. On the other hand, T2 was not interesting in that amplification always eventually leads to saturation of the listening media. That is why an X value is given in table 2. The case where α was set randomly on all channels (within a limited range) revealed that the relative importance of channels can not be strongly altered, in agreement with many speech analysis work. We made an additional test in which the envelope is low-passed. Although low-passing the envelope of a narrow-band signal at half the corresponding filter's bandwidth still ensures perfect reconstruction, our HS does not perceive degradation even in the case where high-frequency channels envelopes are low-passed at 300 Hz. Reduction of this value lead to progressive degradation.

Phase: test T3 . Many experiments showed that phase information contributed little to speech intelligibility [3]. Our findings are in perfect agreement with this as phases shifts of $0 - 2\pi$ were perceptually inaudible in the case where $\delta t\phi$ was constant on all channels. Yet, it is worth to mention that in the case where shifts are done randomly on all channels, a degradation is perceived by listeners, but seems dependent on the utterance being spoken, which makes it very difficult to draw any conclusions.

Instantaneous frequency: test T4 . The T4 scheme provided the most interesting results and stimulated us to go beyond the primary scope of this work. First, let us point out that scaling the phase by a factor is equivalent to scaling the instantaneous frequency by the same factor. Our HS being more sensible in the low-frequency region, the β factor was not set constant for all channels, but rather according to: $\beta_n = \beta \cdot \frac{CF_n}{CF_N}$ where CF_n denotes centre frequency of channel n.

Results of T4 reveal that the range of β over which no perceptual degradation is heard is quite large, both in the positive (stretching of the IF) and the negative (compression of the IF). We believe this result is very important.

In order to see how relevant these findings are, we verified how spectral and temporal properties of speech are modified by T4-like transformations.

3.2. Changes in temporal and spectral properties (T4).

In figure 2 we show the temporal waveforms of the original signal and two modified versions with β=0.45 and β=-0.25. Although their temporal aspects seem equal on the shown temporal scale, figure 3 reveals that on a shorter time scale the differences are clearly observable.

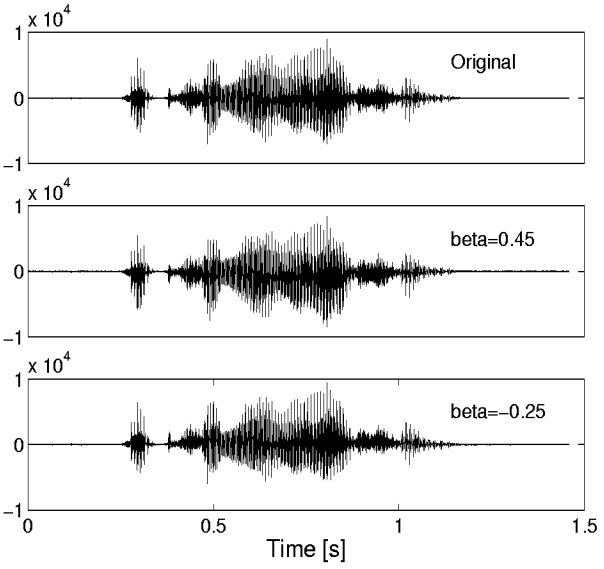

Figure 2: Temporal waveforms of the word *"accumulation"* spoken by and English male.

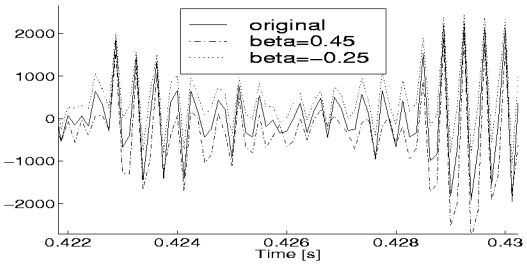

Figure 3: Temporal zoom of the word *"accumulation"*.

Since many speech applications are based on short temporal windows (typically 20 ms), we chose a stationary and a non-stationary segment of the utterance considered here. They are shown at figure 4, and will be used in all further experiments. In figure 5 we show the power spectral densities of the non-statinary parts of the three signals. It appears clearly that the spectra hold major differences on all the frequency axis, even though the β factor affects more the high

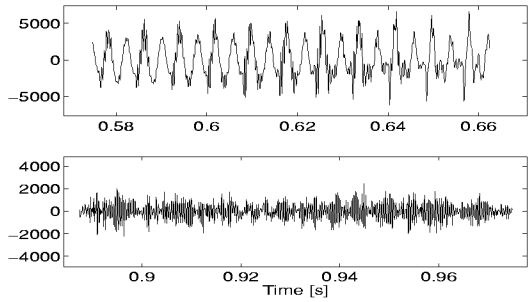

Figure 4: Stationary and non-stationary segments of the word *"accumulation"*.

frequency region (see above). The spectra of the stationary parts are not shown since no major differences are observed between all three signals. This can be explained by the fact that most energy is concentrated in low frequencies, where the β parameter is small. The actual value of β of course influences the differences. These important observations sug-

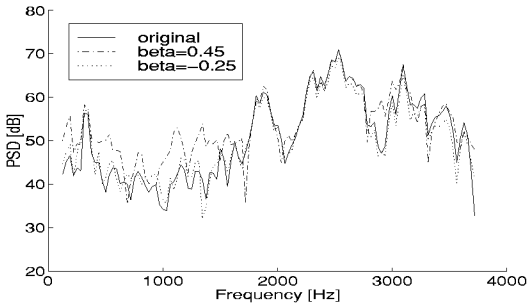

Figure 5: Power spectral densities of non-stationary sample.

gest that speech applications based on the signal's spectra or temporal properties may not be as robust to T4-type transformations than is our HS.

3.3. Effect on analysis methods (T4)

In this section, we consider LPC and cepstral analysis methods and see how their representation of the signal is affected by T4 changes. Coefficients were computed on 25 ms windows. In the LPC case, we used a 12th order predictor (12 coefficients) and in the cepstral case we retained the first 12 coefficients, know to be the most significant [3]. Results are presented in figure 6 for stationary and non-stationary samples of the original and the two modified versions.

The general observation is that indeed there are differences in the coefficient values between the original and modified signals, for both stationary and non-stationary segments. Looking more carefully, three points can be made: (1) the LPCs differ significantly; (2) in both analysis schemes, the non-stationary sample suffers more from these modifications than does the stationary one and (3) cepstral coeffi-

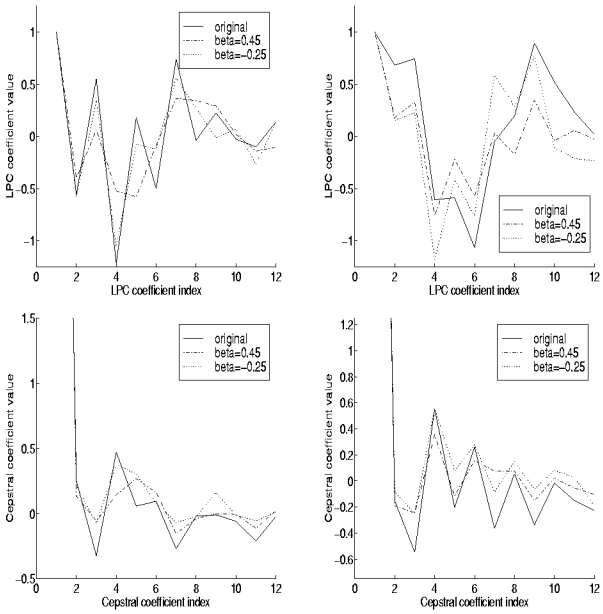

Figure 6: LPC (top) and cepstral (bottom) coefficients for stationary (left) and non-stationary (right) samples.

cients seem more robust, but differences are still observed.

Hence, we can conclude that the standard analysis techniques can be questioned and may not be as invariant as our HS to modifications that can happen in many real-life situations where temporal or spectral properties are altered.

4. DISCUSSION AND CONCLUSIONS

We have presented in this paper a study on to what extent changes on envelope and instantaneous frequency of signals are not perceived by our HS. Our main conclusion is that changes according to T4 scheme (in a given range) are not perceptible but alter considerably temporal and spectral properties of the signal, hence the standard analysis representation, suggesting that the later may be questioned from that point of view.

4.1. Discussion

Envelope. Tests on envelope (T1,T2) only lead to the conclusion that time relevance between sub-signals envelopes is important, which is not new. On the other hand, we also showed that high-CF channel's envelope can be limited to 300Hz without degradation of speech quality. This finding could perhaps be used advantageously in speech coding.

Phase. The results of test T3 are in agreement with many studies on the perceptual unimportance of the phase information for our HS. The novelty reported here is the study of random phases changes on each channels but conclusions are hard to draw since perception is affected depending on the utterance being spoken.

Instantaneous frequency. Test T4 lead to the most interesting results. The range of the β scaling factor over which no perceptual differences are noted is large. The HS robustness to IF changes is not surprising if we consider that speech is always constrained to modifications due to our production mechanism, use of different microphones, etc.

We then looked at how spectral and temporal standard analysis techniques are invariant to the same modifications. Results show lack of robustness of the methods. In order to push these conclusions any further, we need to make experiments in which we would inspect how some modifications, unperceived by our HS but present in many real-life situations, may affect the whole chain of speech processing in applications using LPC/cepstral analysis such as speech recognition. For example we could see if the ability to separate classes is affected by such transformations.

4.2. Conclusions

possible applications. Simply by changing the β parameter, one can enlarge artificially a database. Artificial generation of samples or enlargement of databases can include both samples that are perceptually undifferentiable but have different temporal properties (to increase learning databases, etc.) or intentionally include samples with audible degradation that simulate some real-life sources of 'degradation' such as stress, use of microphones, etc. As a concrete example, we are presently testing this on speaker recognition where learning process could be improved by extra data.

Finally, as a conclusion, we feel perceptual invariants to transformation on speech can provide new research directions in speech processing and some reported results could be applied directly in some speech applications.

5. REFERENCES

[1] B. Delgutte. Auditory neural processing of speech. In W. Hardcastle and J. Laver, editors, *The Handbook of Phonetic Sciences*. Blackwell (Oxford), 1995.

[2] P. Maragos, T. Quatieri, and J. Kaiser. Speech nonlinearities, modulations and energy operators. In *ICASSP*, 1991.

[3] J. Picone. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81:1215–1247, 1993.

[4] A. Potamianos. *Speech processing applications using an AM-FM demodulation model*. PhD thesis, Harvard University, 1995.

[5] A. Potamianos and P. Maragos. A comparison of the energy operator and the hilbert transform approach to signal and speech demodulation. *Signal Processing*, 37:95–120, 1994.