PSYCHOPHYSICS OF SPEECH ENGINEERING SYSTEMS

H. Hermansky and M. Pavel
Oregon Graduate Institute, Portland, Oregon, USA

ABSTRACT

The paper reviews two engineering techniques, the Perceptual linear predictive (PLP) analysis and the RelAtive SpecTrAl (RASTA) processing, used in automatic speech recognition and describe their consistencies with some properties of human speech perception.

INTRODUCTION

Assuming that speech developed so that its linguistically important components are heard well, processing of speech should respect properties of human hearing. However, a blind copying of nature without deeper understanding of underlying mechanisms in hopes of "obtaining" a successful engineering solution has frequently proven to be a failure.

We believe that engineering disciplines can benefit from selective modelling of relevant characteristics of human information processing. In this paper we discuss two techniques, the Perceptual linear predictive (PLP) analysis, and the RelAtive SpecTrAl (RASTA) processing, which were designed to improve the performance of automatic speech recognizers. Subsequently, these techniques were found to be consistent with specific properties of human speech perception. We discuss (in italics) relevant properties of human speech perception, and before describing the two we attempt to put both techniques into historical perspective with selected engineering systems.

PERCEPTUAL LINEAR PREDICTION (PLP)

The PLP analysis technique was designed to suppress speaker dependent components in features used for automatic speech recognition. Several basic properties of human hearing (as noted below, each previously used in engineering) were integrated in a speech analysis technique called PLP [1].

Root Spectral Compression

Perception of intensity appears to be consistent with a compressive type of nonlinearity. In particular, perceived loudness of a steady sound is approximately proportional to a cube root of its power [2].

Lim [3] investigated the use of different compressive functions in homomorphic analysis of speech. He concluded that the cube root compression was optimal with respect to resulting speech quality of re-synthesised speech.

Hermansky et al. [4] experimented with varying compressive functions in linear predictive analysis and found that when the short-term power spectrum of speech is compressed through cube root function, the analysis is the least affected by the fine spectral structure of voiced speech. The root spectral compression also helps in modelling spectral envelope zeros which occur in nasalized and fricative speech sounds.

Furthermore, root compressed power spectrum (root compression with exponents 2-4) appears to be optimal for processing which alleviates additive noise in the acoustic signal (see e.g. [5-7]).

Nonlinear spectral resolution

Decreasing selectivity of human hearing with frequency is one of the best documented and least disputed properties of human auditory perception. Bridle and Brown [8] and later Mermelstein [9], and Davis and Mermelstein [10] proposed to use cosine transform of logarithmic energies (cepstrum) from non-uniformly spaced bandpass filters with bandwidth increasing with frequency. Davis and Mermelstein proposed triangular filters with a shape which is about constant on the mel scale. Mel cepstrum is currently the dominant feature extraction technique in automatic speech recognition.

Nonuniform spectral sensitivity of hearing

For typical levels of human speech under increasing power, the spectral sensitivity of hearing is approximately proportional to a cube root of its power [2].

Lim [3] investigated the use of different compressive functions in homomorphic analysis of speech. He concluded that the cube root compression was optimal with respect to resulting speech quality of re-synthesised speech.

Hermansky et al. [4] experimented with varying compressive functions in linear predictive analysis and found that when the short-term power spectrum of speech is compressed through cube root function, the analysis is the least affected by the fine spectral structure of voiced speech. The root spectral compression also helps in modelling spectral envelope zeros which occur in nasalized and fricative speech sounds.

Furthermore, root compressed power spectrum (root compression with exponents 2-4) appears to be optimal for processing which alleviates additive noise in the acoustic signal (see e.g. [5-7]).

Nonlinear spectral resolution

Decreasing selectivity of human hearing with frequency is one of the best documented and least disputed properties of human auditory perception. Bridle and Brown [8] and later Mermelstein [9], and Davis and Mermelstein [10] proposed to use cosine transform of logarithmic energies (cepstrum) from non-uniformly spaced bandpass filters with bandwidth increasing with frequency. Davis and Mermelstein proposed triangular filters with a shape which is about constant on the mel scale. Mel cepstrum is currently the dominant feature extraction technique in automatic speech recognition.

Nonuniform spectral sensitivity of hearing

For typical levels of human speech under increasing power, the spectral sensitivity of hearing is approximately proportional to a cube root of its power [2].

Lim [3] investigated the use of different compressive functions in homomorphic analysis of speech. He concluded that the cube root compression was optimal with respect to resulting speech quality of re-synthesised speech.

Hermansky et al. [4] experimented with varying compressive functions in linear predictive analysis and found that when the short-term power spectrum of speech is compressed through cube root function, the analysis is the least affected by the fine spectral structure of voiced speech. The root spectral compression also helps in modelling spectral envelope zeros which occur in nasalized and fricative speech sounds.

Furthermore, root compressed power spectrum (root compression with exponents 2-4) appears to be optimal for processing which alleviates additive noise in the acoustic signal (see e.g. [5-7]).
The 5th order PLP model was used successfully in speaker-independent recognition of digits [1]. For more complex tasks with a sufficient amount of training data, higher model order (7th-8th) appears to be more efficient.

RELATIVE SPECTRAL (RASTA) PROCESSING

We will next describe our engineering approach based on certain temporal properties of human hearing.

Perception of modulated signals

Since early experiments of Riesz [16] it is known that sensitivity of human hearing to both the amplitude and the frequency modulation is highest for frequency of modulation at about 4-6 Hz. Thus, human hearing in perception of modulated signals acts as a band-pass filter.

Drullman et al. [17, 18] support the band-pass character of human hearing in speech perception by showing that low-pass filtering of 1/4 octave-derived spectral envelopes of speech at frequencies higher than 16 Hz or high-pass filtering it at frequencies lower than 2 Hz causes almost no reduction in speech intelligibility. They proposed that the bulk of linguistic information is contained in modulation frequencies between 2 and 16 Hz.

Furui [19] introduced delta features to enhance dynamic components of speech signal. To approximate the derivative of time trajectories of cepstral coefficients, Furui computed the delta features using a regression fit to a short segment of the cepstral trajectories. This operation is equivalent to band-pass filtering of the trajectory by an FIR filter with a relatively shallow (-6db/oct) low frequency slope. The optimal length of the segment for deriving the regression fit was about 170 ms, which corresponds to a FIR bandpass filter with its maximum at about 4 Hz [20].

Rosenberg et al. [21] experimented with cepstral mean subtraction in speaker recognition system using mean computed over short-term window of variable lengths. They reported the best results for speech window of 165 ms. As discussed e.g. in Hermansky and Morgan [20], the cepstral mean subtraction with the 165 ms window implies high-pass filtering with the filter cut-off frequency of about 1 Hz.

The Technique

RASTA engineering technique uses the fact that linear distortions and additive noise in speech signal show as a bias in the short-term spectral parameters. Since rate of such extra-linguistic changes is often outside the typical rate of change of linguistic components, Hermansky et al. [22] and Hirsh et al. [23] have proposed filtering of temporal trajectories of speech parameters which would alleviate the extra-linguistic spectral components from the speech representation. This technique is known as RASTA speech processing. A series of recognition experiments in which the test data were linearly distorted by convolution with a simple first-order high-pass system [20] was run with different RASTA filters to determine the optimal filter structure. Results of experiments are shown in Fig. 3.

The optimal filter for recognition of noisy speech was found to be a bandpass filter with the pass-band between about 1 Hz and 12 Hz. The time constant of the integrator in the filter was about 170 ms. RASTA processing enhances dynamic events in the signal and suppresses the slowly varying ones, as illustrated in Fig. 4.
The fifth order PLP analysis of 18 synthetic cardinal vowels yields results which agree well with Bladon and Fant's [27] perceptual experiments: the second spectral peak approximates well the effective second formant F2' [1]. Moreover, the bandwidths of the PLP model preserve information about spread of the underlying formant clusters, thus alleviating a fundamental objection [28, 29] to the F2' concept (see [1] for evidence and discussion). The two peaks of the fifth order PLP model start merging when their distance approaches 3.5 Bark, thus being consistent with [13].

Hermansky and Broad [30] demonstrate a high correlation between positions of the second spectral peak of the fifth order PLP model and the resonance frequency of the uncoupled front cavity of the simulated vocal tract of front and mid vowels, used in articulatory synthesis of the vowel-like sounds. Table I is a summary of their results. The first row contains correlations of the tract length and the resonance frequency of the uncoupled front cavity with the second peak of PLP model, extracted from the synthesized speech. The second row shows averaged correlations with the first four formants. Note that the formant frequencies, which are strongly dependent on anatomy of the particular vocal tract, correlate highly with the tract length. The weak correlation of the second peak of the PLP model with the tract length implies its relative independence of the talker. Its strong correlation with the resonance frequency of the uncoupled front cavity supports Fant's proposal of its correspondence with the effective second formant F2' [26].

<table>
<thead>
<tr>
<th>Second Peak of PLP Model</th>
<th>Tract Length</th>
<th>Front cavity resonance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formants (Averaged)</td>
<td>-0.71</td>
<td>0.22</td>
</tr>
</tbody>
</table>

The results of Hermansky and Broad [30] and Fant and Risberg [25] are in good agreement. Just as more or less accepted that the formants are extracted by some form of integration of a fundamental frequency peaks.

Later [31] they also show a high correlation of the PLP-estimated F2' with the front cavity resonance estimated from the x-ray microbeam data. Additional work is needed to get full support for their hypothesis.

**RASTA and forward masking**

If a loud sound is followed closely in time by a weaker sound, the audibility of the weaker sound is diminished. This effect, called forward masking, reflects a significant nonlinearity since, independently of the masker amplitude, the effect seems to last for about 200 ms (see e.g. [32]).

As we noted earlier, the phenomenon of forward masking reflects aspects of temporal properties of the auditory system. Forward masking effect is typically measured by presenting, on each trial, a masker (tone or band-passed noise) for 200 milliseconds or longer. Human observers are asked to detect a brief probe presented after a variable delay following the offset of the masker. The masking effect is summarised by the size of the auditory system. For short delays, the masking effect is determined by the masker level. However, the masking effect decays rapidly, and becomes negligible for delays greater than 200 milliseconds, independent of the masker level. The decaying dependence of the masking effect on the logarithmic delay is well approximated by a set of straight lines that intersect at a point corresponding to the delay of approximately 200 milliseconds. This is illustrated in Fig. 6 by the shaded triangle which was derived from human data for 1kHz and 30-60 dB SPL maskers (experiment 1 in [34]).

Prior attempts to account for the data led researchers to models based on automatic gain control such as proposed by [32]. In his model, the effect of the masker was to reduce temporarily the system gain. Although this model could account for the temporal behaviour of forward masking data, it did not specify a plausible process for the temporal dependency of the gain.

A decade later, a scrutiny of the RASTA engineering model provided two interesting insights [33]. First, a reduction in gain in the AGC model is equivalent to a subtraction preceded by a logarithmic transformation. Second, exponential decay in the logarithmic domain with appropriate choices of time constant can produce data that closely approximate linear decay. Both such operations are implemented in the RASTA model.

To investigate the potential of RASTA processing for modeling the temporal masking effect, we duplicated a part of experiment 1 from [34]. Critical-band spectra were computed by PLP analysis using 1 kHz stimuli. The critical-band spectra were processed by our standard RASTA filter [20]. Probe detection was mediated by a comparison of a spectral distance measure of RASTA processed loudness profiles (critical-band spectra in cube-root power) of a masker alone and of the masker followed by a probe. The process is illustrated in Fig. 5.

---

Fig. 5 A model of the experiment for investigation of temporal properties of RASTA processing.

Results, shown in Fig.6, are qualitatively consistent with conclusions from human forward masking experiments [34] with implications indicated in the figure by the shaded triangle overlaid over our data. To obtain the fit, we allowed for a linear
optimization of the distance measure, i.e. the actual Euclidean distance between loudness profiles was multiplied by a constant. (0.12) and another small constant (0.9) was added to the result.

ACKNOWLEDGEMENTS

This work has been supported in part by NSF-ARPA Grant 1RI-9314959 to Oregon Graduate Institute. The authors thank Steven Greenberg for pointing out the relation of RASTA processing to perception of modulation and for useful editorial suggestions.

REFERENCES


[34] Jesteadt, W., Sid P. Bacon, and James R. Lehman, Forward masking as a function of frequency, masker level, and signal delay. J. Acadus. Soc. Am., 1982,