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[1].

We believe that engineeringdisciplines can benefit from *selective* modelling of *relevant* characteristics of human information processing[2]. In this paper we discuss two techniques, the Perceptual linear predictive (PLP) analysis, and the RelAtive SpecTrAl (RASTA) processing, which were designed to improve 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.

[1] Airplanes do not flap wings, and most of "auditory models" do not demonstrate significant advantage in engineering, and sometimes yield clearly inferior results.

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 bellow, 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])

[2] Airplanes do not flap their wings, but their design is based on thorough understanding and use of principles of aerodynamics which allow birds to fly.

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 communication, hearing is most sensitive in 2-4 kHz range, therefore emphasising the second and third formant region.

A typical preemphasis in speech analysis approximates this property by 6dB/oct high-pass filtering of the signal.

To obtain more stable formant estimates, Itahashi and Yokoyama [11] proposed to warp the spectral envelope of speech (estimated by high-order LPC analysis) using a mel warping function, and weight it by an approximation of the Fletcher-Munson equal loudness curve. The resulting auditory-like spectrum was then again approximated by relatively low (6th order) LPC all-pole model.

Broad spectral integration in speech perception.

Klatt [12] *speculated that for gender normalization, larger than 1 Bark spectral resolution would be required. This notion is supported by perceptual studies that suggest that human speech perception could integrate formant peaks within 3.5 Bark interval* [13], *and therefore could merge several speech formants. Thus, frequency resolution for perception of speech signals seems to be considerably broader that the critical-band concept would suggest*

Pols et al. [14] reported that the first three (six) principal components of a set of non-uniformly spaced 1/3 octave filter bank output power explain 82% (97%) of variance in his data. Later work Pols [15] also shows that these first three principal components can be used successfully in automatic speech recognition.

The Technique

Several engineering approximationsto the properties of human speech perception are used in PLP analysis of speech:

1) critical band (Bark) nonlinear frequency resolution, implemented by integrating short-term Fourier spectrum of speech under increasingly wider trapezoidal curves,

2) asymmetries of auditory filters, implemented by relatively steep (25dB/Bark) slope of the trapezoidal curve towards higher frequencies and more gradual (10dB/Bark) slope towards lower ones,

3) unequal sensitivity of human hearing at different frequencies, implemented by a fixed approximation of Fletcher-Munson equal loudness curve,

4) intensity-loudness nonlinear relation, implemented by a cube root compression, and

5) broader than critical-band integration, hypothesised in perception of speech (see e.g. [12]), implemented by an autoregressive all-pole model.

The optimal order of the PLP all-pole model was determined experimentally on cross-speaker speech recognition experiments in which training data from one speaker were used to recognise speech of another speaker. Results are shown in Fig. 1. The two-peak (5th order) model was found to be optimal.

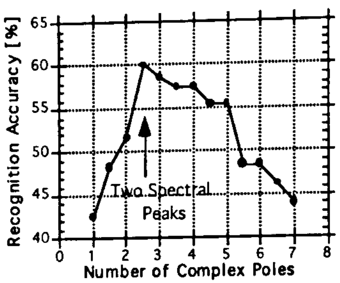

Fig. 1 Dependency of recognition accuracy in cross-speaker experiment on the maximum number of spectral peaks (model order) of PLP model

The spectrograms in Fig. 2 show that, in comparison to the conventional formant based representations, the broader spectral integration implied by low-order PLP analysis is capable of more consistent speech representations from adult and child speech.

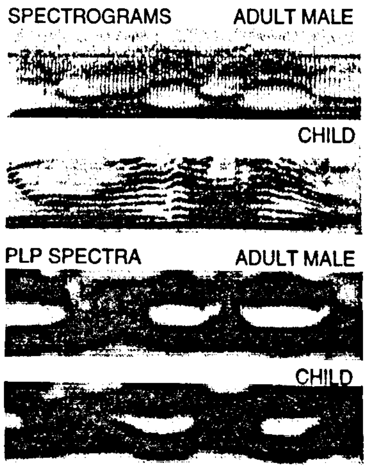

Fig. 2 Spectra of adult and child speech obtained by conventional spectral analysis and by PLP analysis

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 that 2 Hz causes almost no reduction in speech intelligibility. They proposed that 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 regressionfit 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 the 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.

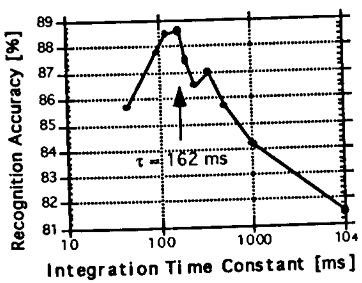

Fig. 3 Dependency of recognition accuracy in presence oflinear distortions on time constant of integrator of RASTA filter.

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 is the signal and suppresses the slowly varying ones, as illustrated in Fig. 4.

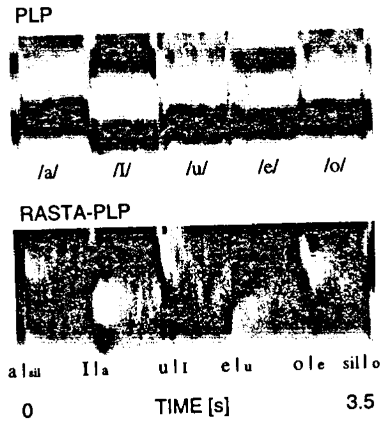

Fig. 4 Spectra of five sustained Czech vowels obtained by PLP and RASTA-PLP analyses. Note enhanced transitions resulting from RASTA processing

The RASTA band-pass filtering is typically done either on logarithmic spectrum (or cepstrum, which is a linearly transformed logarithmic spectrum) or on the spectrum compressed by *ln(const+x)* nonlinearity. However, Hermansky et al. [7] reported that RASTA filtering on root-compressed power spectrum (with filters designed from the training data) is effective for perceptual enhancement of noisy telephone speech. Filters in the frequency range with most speech energy have a maximum at about 6-8 Hz.

For speech recognition applications, we most often use RASTA processing in combination with the above described PLP technique. In this combination, RASTA filtering is performed on outputs from a critical-band analysis, i.e., prior to the cube root compression and loudness equalisation, and the all-pole modelling. We note that RASTA-PLP technique is rapidly gaining recognition in engineering community, especially in applications which can tolerate or even benefit from the enhanced spectral dynamics, such as the isolated phrase recognition. Alternative recognition paradigms which could capitalise on the enhanced spectral dynamics are being studied for applications of RASTA processing in recognition of continuous speech [24].

CONSISTENCIES WITH HUMAN SPEECH PERCEPTION

Although both PLP and RASTA were designed on purely engineering grounds and with a clear engineering objective in mind, they both turned out to be at least in certain aspects consistent with human speech perception.

PLP and effective perceptual second formant F2'

Fant and Risberg [25] *observed that all Swedish vowels can be simulated by synthetic stimuli with only two spectral peaks, providing that their second spectral peak F2' is in particular position, which does not necessarily coincide with any of the formants. Fant* [26] *proposes that the effective second formant F2' might correspond to a resonance frequency of the uncoupled front cavity of the vocal tract. Hermansky and Broad* [30] *showed on X-ray tracings that the front cavity appears to be less dependend on the age of the talker than the rest of the vocal tract. They speculated that speech perception (simulated by the PLP analysis) might be able to integrate detailed formant structure and extract the resonance frequency of the front part of the vocal tract.*[3]

The 5th 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 5th 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 5th 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 1. is a summary of their results. The first row contains correlations of the tract legnth 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 1.

	Tract Length	Front cavity resonance
Second Peak of PLP Model	-0.18	0.9
Formants (Averaged)	-0.71	0.22

[3] Just as is 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 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 sound level of the probe, above its threshold, required for fixed detection performance.

Typical data from such experiments exhibit features that implicate non-linear aspects 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 extrapolated mean 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.

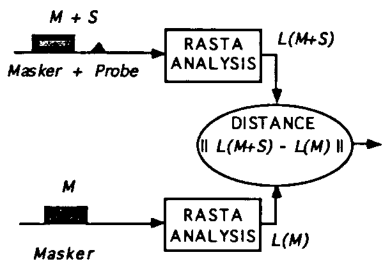

Fig. 5 A model of the experiment for investigation of temporal properties of RASTA processing.

Results, shown in Fig. 6, are qualitatively consistent with conclusions from human forward masking experiments [34] which implications are 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.

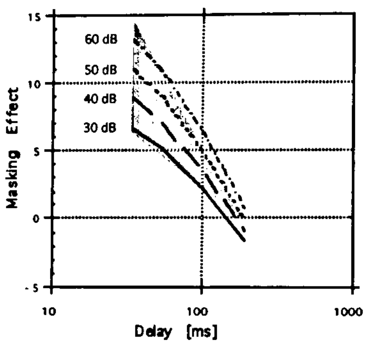

Fig. 6. Spectral distances betweena tonal masker and the masker with a probe. The parameter is the level of the masker. Extrapolated human performance [34] is shown by shaded area.

CONCLUSIONS

We have reviewed two successful engineering approaches, designed to alleviate sensitivity of speech processing to extra-linguistic factors and used widely in speech engineering. We noted the similarity in their behaviour to that of the human auditory system. Some of these consistencies were obtained because of an explicit motivation to model the human auditory system, but others were strictly the results of engineering optimisations.

We have also noted that analysis of engineering systems may lead to new insights into the processes underlying human auditory perception. There are instances where engineering technique, even though designed only as a practical solution to a particular engineering problem, turned out to be a good model of human auditory perception.

ACKNOWLEDGEMENTS

This work has been supported in part by NSF-ARPA Grant IRI-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

[1] Hermansky, H., *Perceptual linear predictive (PLP) analysis of speech.* J. Acoust. Soc. Am., 1990. **87**(4): p. 1738-1752.

[2] Stevens, S.S., *On the psychophysical law.* Psychol. Rev., 1957. **64**: p. 153-181

[3] Lim, J.S., *Spectral root homomorphic deconvolution system.* Proc. IEEE ASSP-27, 1979. **27**(3): p. 223-233.

[4] Hermansky, H., H. Fujisaki, and Y. Sato. *Analysis and synthesis of speech based on spectral transform linear predictive method.* in *Int. Conf. Acoust. Speech and Sig. Proc.* 1983.

[5] Porter, J.E. and S.F. Boll. *Optimal estimators for spectral restoration of noisy speech.* in *Int. Conf. Acoust. Speech and Sig. Proc. 84*. 1984.

[6] Hanson, B. and D. Wong. *The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence in interfering speech.* in *Int. Conf. Acoust. Speech and Sig. Proc.* 1984.

[7] Hermansky, H., E. Wan, and C. Avendano. *Speech enhancement based on temporal processing.* in *Int. Conf. Acoust. Speech and Sig. Proc.* 1995.

[8] Bridle, J.S. and M.D. Brown. *An experimental automatic word recognition system.* in *JSRU Report No. 1003*. 1974. Ruislip, England: Joint Speech Research Unit.

[9] Mermelstein, P., *Distance measures for speech recognition, psychologicaland instrumental*, in *Pattern Recognition and Artificial Intelligence*, R.C.H. Chen, Editor. 1976, Academic Press: New York. p. 374-388.

[10] Davis, S.B. and P. Mermelstein, *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.* Trans. IEEE ASSP-28, 1980(4): p. 357-366.

[11] Itahashi, S. and S. Yokoyama. *Automatic formant extraction utilizing auditory equal loudness contour,.* in *Fall Meeting Acoust. Soc. Japan*. 1974.

[12] Klatt, D.H., *Speech processing strategies based on auditory models*, in *The representation of speech in the peripheral auditory system*, R. Carlson and B. Granstrom, Editors. 1982, Elsevier - Biomedical Press: New York. p. 181-202.

[13] Chistovich, L.A., *Central auditory processing of peripheral vowel spectra.* J. Acoust. Soc. Am., 1985. **77**: p. 789-805.

[14] Pols, L.C.W., L.J.T.v.d. Kamp, and R. Plomp, *Perceptual and physical space of vowel sounds.* J. Acoust. Soc. Am., 1969. **46**: p. 458-467.

[15] Pols, L.C.W., *Real-time recognition of spoken words.* IEEE Trans. Computers, 1971. **20**: p. 972-978.

[16] Riesz, R.R., *Differential intensity sensitivity of the ear for pure tones.* Phys. Rev., 1928. **31**(Ser. 2): p. 867-875.

[17] Drullman, R., J.M. Festen, and R. Plomp, *Effect of temporal envelope smearing on speech reception.* J. Acoust. Soc. Am., 1994. **95**: p. 1053-1064.

[18] Drullman, R., J.M. Festen, and R. Plomp, *Effect of reducing slow temporal modulations on speech reception.* J. Acoust. Soc. Am., 1994b. **95**: p. 2670-2680.

[19] Furui, S., *Cepstral analysis technique for automatic speaker verification.* IEEE ASSP-29, 1981(2): p. 254-266.

[20] Hermansky, H. and N. Morgan, *RASTA processing of speech.* Proc. IEEE Speech and Audio Processing, 1994.

[21] Rosenberg, A.E., C. Lee, and F.K. Soong. *Cepstral channel normalization techniques for HMM-based speaker verification.* in *International Conference on Spoken Language Processing*. 1994. Yokohama, Japan.

[22] Hermansky, H., N. Morgan, and P. Kohn. *Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP).* in *Eurospeech '91*. 1991. Genova.

[23] Hirsh, H.G., P. Meyer, and H. Ruehl. *Improved speech recognition using high-pass filtering of subbband envelopes.* in *Eurospeech '91*. 1991. Genova.

[24] Morgan, N., *et al. Stochastic perceptual models of speech.* in *Int. Conf. Acoust. Speech and Sig. Proc. 95*. 1995.

[25] Fant, G. and A. Risberg. *Auditory matching of vowels with two formant synthetic sounds.* in *STL-QPRS 2-3*. 1962. Stockholm: Royal Institute of Technology.

[26] Fant, G., *Acoustic theory of speech production.* 2nd printing ed. 1970, The Hague: Mounton. p.123.

[27] Bladon, A. and G. Fant. *A two-formant model and the cardinal vowels.* in *STL-QPRS*. 1978. Stockholm: Royal Institute of Technology.

[28] Fujimura, O., *On the second spectral peak of front vowels: a perceptual study of the role of the second and third formants.* Language and Speech, 1967: p. 10181-10193.

[29] Bladon, A.W., *Two-formant models of vowel perception: shortcomings and enhancements.* Speech Communication, 1983. **2**: p. 305-313.

[30] Hermansky, H. and D. Broad. *The effective second formant F2' and the vocal tract front cavity.* in *Int. Conf. Acoust. Speech and Sig. Proc.* 1989.

[31] Broad, D. and H. Hermansky, *The front cavity/F2' hypothesis tested by data on tongue movements.* J. Acoust. Soc. Am., 1989. **86**(Suppl. 1): p. S13-S14.

[32] Pavel, M., *Homogeneity in complete and partial masking*, . 1980, New York University.

[33] Pavel, M. and H. Hermansky, *Temporal masking in automatic speech recognition.* J. Acoust. Soc. Am, 1994. **95**(5): p. 2876.

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