AUTOMATIC SPEECH RECOGNITION AND HUMAN AUDITORY PERCEPTION

Hynek Hermansky *

ABSTRACT

Low-order all-pole model of critical-band auditory spectrum, the Perceptually based linear predictive (PLP) model, compared with the standard linear predictive (LP) model, yields higher recognition accuracy in speaker-independent automatic speech recognition (ASR) while offering significant computational savings. Paper reviews the principle of the PLP method, shows and discusses some ASR results, and experimentally evaluates and discusses relationships of the PLP parametric representation of speech to several theories of speech perception. Results support use of model of speech perception in ASR front-ends.

INTRODUCTION

Many speech analysis techniques used in the ASR front-end inherently employ speech production models. The speech production model contains speaker-dependent, linguistically irrelevant cues. The task in ASR is to develop the analysis-metric combination, consistent with human speech perception. Hermansky et al. [6,7,8] demonstrated that modeling some basic psychophysical properties of human auditory perception yields a low-dimensional parametric speech representation, suitable for use in the current ASR. This representation, which we call the PLP representation, is shown in comparison with the standard LP in the Fig. 1. Fig. 2 shows a comparison of several different ASR front-ends in speaker-independent isolated-digit ASR. The psychophysically motivated low-dimensional PLP representation yields the best result. The advantage of the PLP in ASR has been also observed in [1,11].

By showing consistency of the PLP speech representation with several theories of speech perception, the present paper argues that the performance of current ASR systems can be significantly improved by a model of human speech perception in the ASR front-end.

Fig.1 Spectra of all-pole models approximating the speech power spectrum (LP 14th order) and the speech auditory spectrum (PLP 5th order). LP analysis estimates formants of speech, PLP analysis estimates two most dominant clusters of spectral energy. Displayed word is female "nine". Note different nomination of axis.

Fig.2 Result of isolated-digit speaker-independent ASR. 48 male and female speakers as a test, another 48 male and female speakers in training. DTW template matching, 2 templates/word, result averaged over 48 different template combinations. Savings in the recognition cost are due to fewer parameters in speech representation.

* Speech Technology Laboratory, 3888 State Street, Santa Barbara, California 93105, U.S.A.
**PLP TECHNIQUE**

- Fourier Transform
- Magnitude Squared
- Critical Band Integration
- Equal Loudness Preemphasis
- Intensity-to-Loudness Compression
- Inverse Fourier Transform
- Linear Equation Solving
- Parameter Transform

### Approximate Cost of Analysis (in Numbers of Multiplications)

<table>
<thead>
<tr>
<th>Method</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard LP14</td>
<td>200</td>
</tr>
<tr>
<td>Window</td>
<td>200</td>
</tr>
<tr>
<td>Autocorrel</td>
<td>2000</td>
</tr>
<tr>
<td>Domain</td>
<td>200</td>
</tr>
<tr>
<td>Cepstrum</td>
<td>200</td>
</tr>
<tr>
<td>TOTAL</td>
<td>3900</td>
</tr>
<tr>
<td>PLP</td>
<td>200</td>
</tr>
<tr>
<td>Window</td>
<td>200</td>
</tr>
<tr>
<td>FFT</td>
<td>2000</td>
</tr>
<tr>
<td>Crit.Band</td>
<td>450</td>
</tr>
<tr>
<td>Crit.Root</td>
<td>150</td>
</tr>
<tr>
<td>Inverse FFT</td>
<td>50</td>
</tr>
<tr>
<td>Domain</td>
<td>50</td>
</tr>
<tr>
<td>Cepstrum</td>
<td>50</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1000</td>
</tr>
</tbody>
</table>

1. Limited-Band Integration (100 Bark)
2. Look-Up Table 4 Interpolation

**Fig. 3** Block diagram of the PLP speech analysis method. Estimated computational cost of standard LP and PLP analyses also shown.

**SOME PROPERTIES OF PLP BASED ASR FRONT-END**

<table>
<thead>
<tr>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>±10%</td>
<td>±10%</td>
<td>±10%</td>
<td>±50%</td>
<td>±50%</td>
<td>±50%</td>
<td>±5 dB/oct.</td>
</tr>
</tbody>
</table>

- LP14-CEP
- PLP5-RPS

**Fig. 4** Sensitivity of the standard LP based ASR front-end with cepstral metric and the PLP based front-end with RPS metric to changes in some production parameters of synthetic speech. Area of each symbol proportional to the distance between the reference frame and the frame with modified speech production parameter, averaged over nine synthetic vowel-like sounds which cover the vowel formant space.

The PLP technique represents several properties of human auditory perception: 1) relatively low spectral resolution, decreasing with frequency; 2) low sensitivity at low frequencies; 3) non-linear relation between intensity of sound and loudness. PLP is computationally as efficient as standard LP. As in standard LP, the result is an all-pole function. The PLP spectrum approximates only the major peaks of the critical-band auditory spectrum. As discussed in [7,8], the efficient metric for PLP vectors is the Root power sum (RPS) metric [14], which approximates difference in spectral derivatives of the phase spectrum (group-delay spectra) [10]. Block diagram of the PLP method is shown in Fig. 3. See [6,7,8] for details.

**EFFECT OF THE MODEL ORDER**

- In PLP model order beyond a certain point decreases recognition accuracy. This effect has been observed in several experiments. One experiment dealt with cross-speaker recognition of single-frame, pheme-like spectra (see Fig. 5). The other experiment dealt with speaker-trained and cross-speaker ASR of isolated words (see Fig. 6). In cross-speaker ASR (where only the speaker-independent information contributes to recognition) recognition accuracy is highest for the 2-peak model (4 or 5th order model). Details of the experiments can be found in [8]. The experiment if Fig. 6 confirms earlier results, speakers were different from [8].

**Fig. 5** Averaged recognition accuracy (top 3 candidates) in the recognition of single-frame phoneme-like spectra of two male and two female speakers. 41 phoneme-like spectra for each speaker obtained as centroids of spectral clusters, extracted by hand labeling speech data (500 words/speaker) in significant phonemic points.

**Fig. 6** Averaged recognition accuracy (top 1 candidate) in the speaker-trained ASR (templates from the same speaker as test) and the cross-speaker ASR (templates from single speaker, different from test). Aplha numeric vocabulary (86 words), two male and two female speakers, five realization of the vocabulary each.
EFFECTIVE PERCEPTUAL SECOND FORMANT F2'

Highest recognition accuracy in cross-speaker ASR is obtained with the two-peak spectral model. The two-peak model was proposed for vowel representation by Fant [3]. Experimentally determined F2' frequencies of Fant's model are available for the set of synthetic cardinal vowels [2]. As seen in Fig. 7, positions of spectral peaks of the 5th order PLP model agree rather well with the experimentally observed frequencies of the second effective formant F2' (See note 1).

Unlike the Fant's F1, F2' model, our 5th order PLP model has not two, but five degrees of freedom. The first two degrees of freedom are, (consistently with the Fant's model), positions of spectral peaks. Additional degrees of freedom are bandwidths of PLP spectral peaks, (i.e. measures of spectral spread of the original formant clusters) and the overall spectral slope of the auditory spectrum. Fujimura [4] and Bladon [2] argued against the F1, F2' model by constructing perceptually distinct vowel pairs with identical F2'. Fig. 8 shows that the 5th order PLP model discriminates between such perceptually distinct vowel pairs. The RPS distances agree well with Bladon's perceptual evaluations [2]. We obtained similar result also for Fujimura's ambiguous vowels.

SPECTRAL PEAK INTEGRATION IN THE PLP MODEL

As seen in the Fig. 7, the two-peak PLP model merges into the one-peak model for all vowels which have their F1, F2' peaks closer than 3.0 to 4.0 Barks (see note 2). There is no explicit mechanism in our PLP model for this peak integration. It is a consequence of the low order of the model. We optimized the model order by ASR experiments. (See also note 3). Yet, the range of the spectral peak integration for the 5th order PLP model agrees with spectral peak integration observed in perceptual experiments with speech-like stimuli of Chistovich et al. [5]

1) We have shown earlier for several natural American vowels [6], that the positions of peaks of 5th order PLP model are consistent with F1, F2' frequencies computed from vowel formant frequencies using the Carlson's formula [4].
2) The 3.5 Bark spectral peak integration has been also observed earlier on the /a/-/æ/ synthetic speech continuum [6].
3) In their ASR experiments with variable spectral resolution metrics, Itakura and Umezaki [10] observed the best ASR performance for a metric with rather broad 300 Hz spectral resolution (300 Hz = 3 to 4 Barks in the F1,F2 region.)
PLP AND FRONT CAVITY RESONANCE

Kuhn [12] proposed the front-cavity resonance of the vocal tract as the dominant component in coding the linguistic information. Furthermore, he speculated that the front cavity resonance in vowels correlates with the perceptual F2'. In a pilot experiment, we have analyzed voiced and fricative productions of the same sentence by the standard 10th order LP and the 5th order PLP methods. As seen in Fig. 9, the standard LP models of voiced and fricative productions are quite different. The PLP method estimates similar models for both productions. This suggests the PLP model does not reflect all gestures of vocal organs but only some linguistically more relevant ones.

![Fig. 9 Positions of peaks of the standard 10th order LP model (upper part of the figure) and the 5th order PLP model (lower part of the figure) of the voiced production (left part of the figure) and fricative production (right part of the figure) of the sentence "Where were you a year ago?". In the LP model, positions of the main resonance of the fricative production alternate between the second and the third formant of the voiced production. Note similarity of PLP models of both the voiced and the fricative productions.]

CONCLUSION AND DISCUSSION

We have shown that the positions of the second peak of the 5th order PLP model are consistent with the effective second formant frequency F2' of vowels [2,3]. When the peaks of the 5th order PLP model are closer than 3 to 4 Barks, they merge into the single peak. Such a spectral peak integration in speech perception has been observed earlier [5]. Finally, we have indications that the position of the second peak of the 5th order PLP model coincides with the frequency of the purported vocal tract front-cavity resonance, which has been proposed earlier as both the production correlate of the perceptual F2' of vowels and the acoustic cue in stop consonant perception [12,13].

The problem in current ASR may not be how to utilize all available cues in the speech signal but how to selectively suppress most of them. In comparison with the formant representation estimated by the standard LP, the PLP speech representation is simpler. Yet, it yields higher recognition accuracy even on the highly confusible alphanumeric vocabulary. Thus, the standard (formant-based) speech representation might be making the situation in the ASR look more complicated than it really is. Many irrelevant cues in the speech signal which do not contribute to the decoding of linguistic information seem to be suppressed by human speech perception. The process can be simulated reasonably well by a simple engineering model, based on psychophysics of human auditory perception. Such a model can be used with advantage in current ASR front-ends.

References: