Stochastic techniques in deriving perceptual knowledge.

Hynek Hermansky

IDIAP Research Institute, Martigny, Switzerland

Abstract

The paper argues on examples of selected past works that stochastic and knowledge-based approaches to automatic speech recognition do not contradict each other. Frequency resolution of human hearing decreases with increasing frequency. Spectral basis designed for optimal discrimination among different phonemes of speech have similar property. Further, human hearing is most sensitive to modulations with frequency around 4 Hz. Filters on feature trajectories, designed for optimal discrimination among phonemes of speech are bandpass with central frequency around 4 Hz.

1. Introduction

The speech signal originates in a speaker vocal organs, is processed by the human auditory system, and has as its purpose communication among human beings. This knowledge can be used to advantage in designing speech processing techniques. Such techniques are often called knowledge-based processing techniques.

Variability of the speech signal due to non-linguistic sources of information such as environmental noise, or anatomical and idiosyncratic differences among speakers is not well understood and its influence on the signal appears almost random. The so-called stochastic speech processing techniques that attempt to deal with this random component in the signal are currently dominating the field.

When it comes to reduction in the information rate for speech recognition, both the deterministic and stochastic techniques use knowledge. The difference is that in the knowledge–based techniques the knowledge may come from relevant experiments on speech production and speech perception while in the stochastic techniques the knowledge comes from large amounts of training data.

Most would agree that the strategies using knowledge derived from data appear to work better. On the other hand, the stochastic techniques do require large amounts of the data to get the knowledge. One cannot help wondering if the stochastic techniques do not waste the data on re-learning the same speech-specific knowledge every time again and again. Is there any way to use the knowledge derived by the stochastic techniques from one data set on a new problem? What is it that the stochastic techniques derive from the data?

Given that speech evolved to be heard, it should not be surprising that stochastic techniques optimized on large amounts of speech data turn out to be consistent with relevant properties of human speech production and perception.

2. Linear discriminant analysis

Linear discriminant analysis (LDA) is a stochastic technique that attempts to optimise the linear discriminability between classes in the presence of undesirable within-class variability (see e.g. [Hunt 1979, Brown 1987] for some examples of previous use of LDA in ASR). It requires that the class affiliation of each vector in the data used for the analysis be known (i.e. the database must be labelled).

LDA is most often applied to sequences of several short-term feature vectors [Braun 1987]. In such applications, the resulting linear discriminants form two-dimensional filters that are to be applied to the time-feature plane. In this paper, however, we review works that allow for the interpretation of LDA results in terms of either a) variable-resolution spectral bases that may indicate a certain spectral resolution of speech analysis, or b) the FIR RASTA filters that may indicate a

certain range of modulation frequencies, that are desirable for the classification of speech.

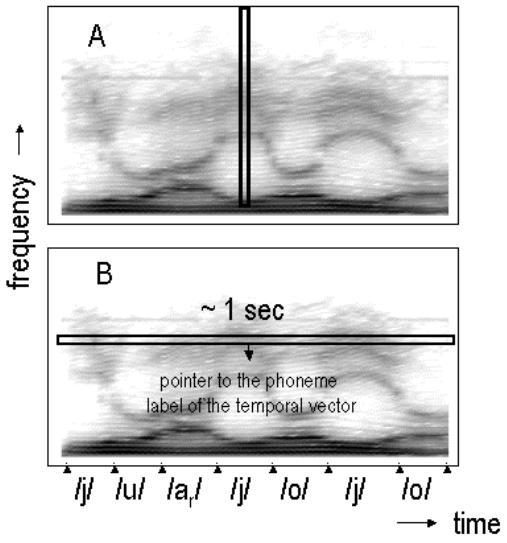

Fig. 1 Two possible ways of forming labelled vectors for LDAT analysis of short-term spectra.

In the first case, reported in [Malayath and Hermansky 2002, 2003] and shown in the A part of Figure 1, LDA was applied to the vector space of logarithmic Fourier spectra of speech. In this case, the resulting LDA discriminant matrix consists of basis for the projection of the short-term spectra. Such basis represent an alternative to the conventional cosine basis of the cepstral projection. The vector labelling is in this case trivial since each vector is clearly affiliated only with a single class. The second way of applying LDA was reported in [van Vuuern and Hermansky 1997] and shown in the B part of Figure 1. In this case, the LDA was applied to the vector space formed from segments of temporal trajectories of spectral energies. The LDA discriminant matrix in this case consists of FIR filters to be applied on time trajectories of logarithmic spectral energies. In the reported experiments temporal vectors about 1 s long were used. Each vector spanned much more than a single phoneme, and was labelled by the phoneme at the centre of the vector.

3. Critical-bands of hearing

Fletcher observed that a signal with frequency components outside a certain "critical band" does not affect the detection of an another signal with its frequency components inside this "critical band" (see [Fletcher 1953] for a review of their earlier experiments that revealed existence of the critical bands). This is clear evidence of the ability of human hearing to separate different spectral components of the acoustic signal into individual bands for further processing. An important property of such "critical-band-like" spectral analysis is that its frequency resolution is lower at higher frequencies [Fletcher 1953]. A similar spectral scale has been observed in experiment with perception of the pitches of tones [Shower and Biddulph 1931]. Even though neither masking experiments nor experiments with the perception of pitch suggest that spectral profiles of sounds are the entities extracted and used by hearing for sound classification, benefits of critical-band-like spectral resolution in ASR is well established through years of comparative ASR experiments. Critical-band-like spectral analysis is often emulated in speech processing by weighted summations of the short-term Fourier spectrum [Davis and Mermelstein 1980, Hermansky 1990].

Malayath and Hermansky [Malayath and Hermansky 2002, 2003] applied LDA to short-term spectral vectors from Fourier analysis (20 ms hamming window, 10 ms analysis step) of all monophthong vowels from the OGI Stories database (OGI Stories contains about 3 hours of fluent American English telephone-quality speech from more than 200 adult speakers of both genders, hand-labelled by phonemes). The LDA was used to find such projections of the logarithmic short-term spectrum (spectral basis) that would allow for optimal discrimination among the vowels.

The first four spectral basis from their LDA analysis are illustrated in Fig. 2. Notice that the period of these spectral basis is shorter at lower frequencies. Subsequently, speech analysis that employs such spectral basis has higher spectral resolution at lower frequencies. [Malayath and Hermansky 2002, Malayath and Hermansky 2003] show by sensitivity analysis of an emulated critical-band filter-bank and the LDA-derived spectral projection that the spectral resolution implied by spectral basis in Fig. 2 is very similar to

the spectral resolution of the auditory-like Bark frequency scale (Fig. 3).

This finding supports earlier results of [Umesh et al. 1997] who derived auditory-like frequency warping by minimizing the differences between speech from different talkers.

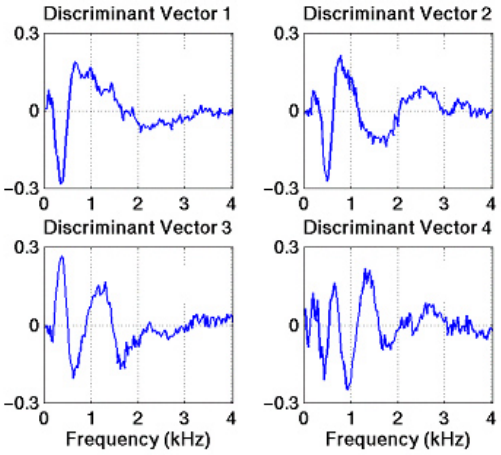

Figure 2 Spectral basis derived by LDA technique

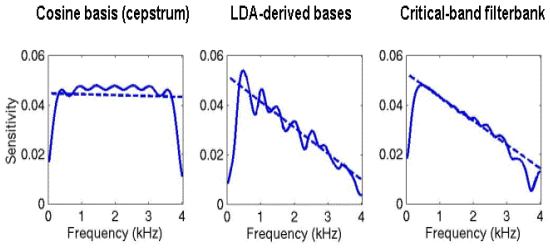

Figure 3 Sensitivity to frequency change of a synthetic formant for three different spectral analysis techniques. As expected, the cepstral projection yields approximately uniform sensitivity while the sensitivity of the LDA-derived projection is similar to the sensitivity of the emulated critical-band filter-bank

The optimality of the logarithmic-like spectral scale in the discrimination of vowels is not surprising. Changes in the position of the tract constriction cause roughly equal *relative* changes in formant frequencies. That is: when the first formant changes from its initial position around 500 Hz by say 10 Hz, the second formant moves from its initial 1500 Hz by about 30 Hz and the third by about 50 Hz) i.e. the resulting vowel spectrum changes about uniformly on the logarithmic frequency scale.

4. Perception of modulations

Since early experiments in perception of modulated signals [Riesz 1928] it has been known that the ear is most sensitive to modulations of around 4 Hz. This finding has been subsequently verified a number of times (see [Kay 1982] for a review). Further, the extensive experiments of Drullman and his colleagues [Drullman et al 1994] and Arai and his colleagues [Arai et al 1999] have shown that only the spectral envelope changes between about 1 and 15 Hz are necessary for maintaining high intelligibility of speech.

For the purpose of modeling, the sensitivity of human hearing to spectral envelope changes can be emulated by filtering temporal trajectories of computed parameters. This has been done in RASTA processing of speech [Hermansky and Morgan 1994] to attenuate features with rates of change that are not expected for speech. The initial ad hoc form of the RASTA filter was optimised on a relatively small series of ASR experiments with noisy telephone digits. The form of the optimised RASTA filter independently confirmed the experiments of Drullman et al and Arai et al. – the filter passed components between 1 and 15 Hz and attenuated slower and faster changing parameters.

Van Vuuren and Hermansky formed a 101-dimensional vector space from logarithmic outputs of an emulated critical-band filter-bank [Hermansky 1990] with vectors labelled by their respective phoneme classes. Each vector then spanned about 1 s at a 100 Hz sampling frequency. LDA analysis yielded a 101 X 101 scatter matrix, decomposed into its principal components. Then the principal vectors were used to represent FIR filters, which most efficiently (with respect to the within-class and the across-class variability) mapped the 101-dimensional input space to a single point of the output space. Since the target classes were context independent phonemes (just as in the previous experiment in the LDA design of

the spectral basis), the FIR filters designed in that way attempted to compensate for a coarticulation with neighbouring phonemes. Further, they also attempted to compensate for other sources of non-linguistic variability such as noise.

Frequency responses of the first three FIR filters derived from about 60 hours of the forcefully-aligned Switchboard database are shown in Fig. 4 from [Hermansky 1998]. Filters for different frequency channels are similar. The frequency characteristic (shown at in the right part of the figure) are generally consistent with RASTA [Hermansky and Morgan 1994], and delta, and double-delta features of speech [Furui 1981]. However, the impulse responses of the data-derived filters shown in the left part of the figure suggest the preference for the zero-phase filters. Effective parts of the impulse responses appear to span at least 250 ms. An interesting fact is that the LDA filters derived at different frequencies (not shown here) are roughly the same, i.e. the filters at, say, 500 Hz do not noticeably differ from the filters derived at 3 kHz. This result would support the notion of the second (post-cochlear) time constant, hypothesized since the early works of Gabor [Gabor 1946].

The general characteristics of the data-derived RASTA filters appear to be relatively independent of the particular database used for their design. The most important processing involves a mild temporal lateral inhibition in which the average of several spectral values around the current time instant is subtracted from the weighted average of spectral values from surrounding past and future contexts. Next is the difference between weighted averages from left and right contexts of the current frame (the first derivative of the first discriminant vector), followed by an aggressive Mexican-hat temporal lateral suppression (the second derivative of the first discriminant vector) implying quite a narrow band-pass filter with a 12dB/oct slope. Such impulse responses can be interpreted as a difference of two Gaussians (the first discriminant) and its derivatives (higher discriminants). Mexican-hat-like dynamics-enhancing functions are hypothesized to be important for scene interpretation by the human visual system [Marr 1982].

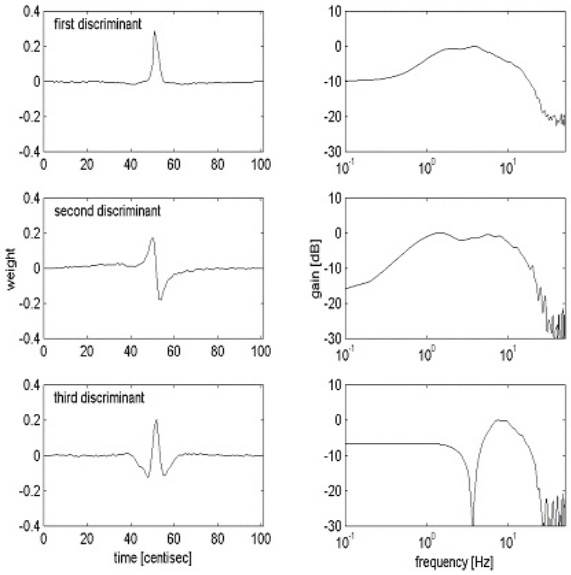

Figure 4 Impulse and frequency responses of the first three discriminant vectors from the LDA-derived discriminant matrix. The filters for the 5 Bark frequency channel are shown here. Filters for the other carrier frequencies studied (between 1 and 14 Bark) are very similar.

5. Data-guided processing and human auditory perception.

Spectral bases derived by LDA shown in Fig. 2 were applied to deriving a small number of linearly-separable features from the short-term FFT logarithmic power spectrum. The only built-in prior knowledge from hearing is the use of the power spectrum. (This may be justified by the frequency selectivity of human cochlea and by the one-way rectification of auditory hair-cell firings). The LDA technique is otherwise rather ignorant in matters of human psychophysics and/or physiology and merely attempts to do the engineering job of efficient separation of speech sound classes. Yet, it delivers spectral resolution that is consistent with human hearing!

The same may be said of LDA-derived RASTA filters. The impulse responses could have been highly concentrated in time but they are not, implying that it is beneficial for the identification of phonemes in running speech to collect data from

relatively large time spans, significantly exceeding the typical 10-20 ms length of the analysis window [Yang et al. 2000]. Rather, consistently with the "critical time interval" observed in forward temporal masking and many other perceptual phenomena, the time span for information extraction is several hundreds of ms. Frequency responses of the dominant discriminants are band-pass, passing the range of modulation frequencies between roughly 1 Hz and 15 Hz, just where human hearing is the most sensitive. Thus, again, the temporal processing that is needed for a good classification of phoneme-like speech sounds is quite consistent with temporal properties of human hearing.

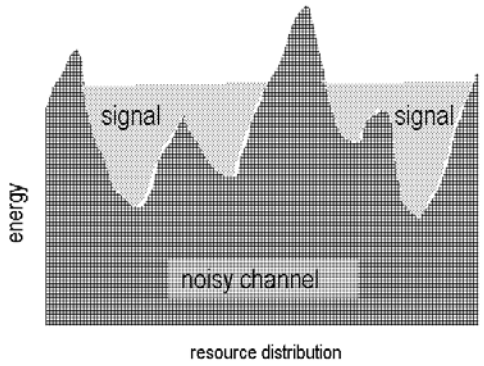

Fig. 5 Optimal distribution of signal energy in noisy channel (adopted from [Galagher 1968])

Why would optimization of a signal processing module on speech data result in human-like processing? Are the properties of human hearing imprinted by speech production mechanism on speech signal? Come to think of it, what else should we expect? Information theory teaches that optimal signal for communication through a noisy channel should confirm to properties of the channel [Gallager 1968]. Then, imprinted on the signal, one should see some properties of the communication channel for which the signal was designed. Hearing likely existed before speech evolved - all parts of the human speech production also serve more life-sustaining function than speech production. It thus appears likely that they were adopted for speaking later in the course of human evolution. Why would not the forces of nature follow the same optimal strategy and form speech to obey the same optimal principles, i.e. to form it in such a way that it is well heard? As a result, when engineer attempts to design an optimal processing strategy, she could end up with the strategy that emulates human hearing!

6. Summary

We discussed speech processing techniques that attempt to optimise processing in such a way that the goal of the processing, i.e. the extraction of information from the speech signal, is better achieved. Such techniques form a bridge between signal processing and stochastic pattern classification and subsequently are trained on large amounts of speech data. The consistency of resulting signal processing modules with some basic properties of human hearing support the notion of speech production evolving to best match the capabilities of hearing.

Acknowledgements

The writing of this paper was supported by DARPA under its EARS program R16007-01, by the Swiss National Science Foundation through the National Centre of Competence in Research on Interactive Multimodal Information Management and by European Community Grants M4 and AMI.

References

T. Arai, M. Pavel, H. Hermansky, and C. Avendano, Syllable Intelligibility for Temporally-Filtered LPC Cepstral Trajectories, J. Acoust. Soc. Am., (105), 5, pp. 2783-2791, May 1999

Brown, P. (1987), The Acoustic-Modeling Problem in Automatic Speech Recognition, PhD Thesis, Computer Science Department, Carnegie Mellon University.

Davis, S.B. and P. Mermelstein (1980), Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans

Drullman, R., Festen, J.M. and Plomp, R. (1994), Effect of reducing slow temporal modulations on speech perception, J. Acoust. Soc. Am., 95, pp. 2670-2680, 1994.

Fletcher, H. (1953), Speech and Hearing in Communication, New York: Krieger.

Furui, S. (1981), Cepstral analysis technique for automatic speaker verification, IEEE Trans. on Acoustic, Speech, & Signal Processing, vol. 29, pp.254-272.

Gabor, D. (1946) Theory of communication, Proc. Inst. Electr. Eng. 93, pp. 429-457, 1946

Gallager, R.G. (1968), Information theory and reliable communication, New York, Wiley.

Hermansky, H. (1990), Perceptual linear predictive (PLP) analysis of speech, J. Acoust. Soc. Am., vol. 87, no. 4, pp. 1738-1752.

Hermansky, H. & N. Morgan (1994), RASTA processing of speech, IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4 pp. 578-589
.
Hermansky, H. (1998) Modulation spectrum in LVCSR, in Research Notes No. 30, 1998, Center for Language and Speech Research, Johns Hopkins University

Hunt, M.J. (1979), A statistical approach to metrics for word and syllable recognition, J. Acoust. Soc. Am., 66(S1), S35(A).

Kay, R.H. (1982) Hearing of Modulations in Sounds, Psychological Review 62 (3), pp. 894-975, 1982.

Malayath, N. and H. Hermansky (2002), Bark resolution from speech data, Proceedings International Conference on Spoken Language Processing 2002, Denver, Colorado, September 2002.

Malayath, N. and H. Hermansky (2003), Data-driven spectral basis functions for automatic speech recognition, Speech Communication, Vol. 40 (4), pp. 446-466, June 2003.

Marr, D. (1982), Vision, W.H. Freeman, San Francisco.

Mermelstein, P. (1976), Distance measures for speech recognition, psychological and instrumental, in Pattern Recognition and Artificial Intelligence, R.C.H. Chen, ed., Academic Press: New York, pp. 374-388.

Riesz, R.R. (1928) Differential intensity sensitivity of the ear for pure tones, Physical Review 31, pp. 867-875.

Shower, E.G. and R. Biddulph, (1931) Differential pitch sensitivity of the ear, J. Acoust. Soc. Am. 3, pp. 275-287.

Umesh, S., L. Cohen and D. Nelson (1997), Frequency warping and speaker normalization, Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Munich, Germany, pp. 983-987.

van Vuuren, S. and H. Hermansky (1997), Data-driven design of RASTA-like filters, Proc. Eurospeech 97, Rhodos , Greece, pp. 409-412.

H. H. Yang and S. Sharma and S. van Vuuren and H. Hermansky (2000), "Relevance of Time-Frequency Features for Phonetic and Speaker-Channel Classification", in Speech Communication (31), pp. 35-50.