ELSEVIER

Speech Communication 25 (1998) 3–27

SPEECH COMMUNICATION

Should recognizers have ears?

Hynek Hermansky [a,b,c,*]

[a] *Oregon Graduate Institute of Science & Technology, Portland, OR, USA*
[b] *International Computer Science Institute, Berkeley, California, USA*
[c] *Technical University, Brno, Czech Republic*

Received 1 September 1997; received in revised form 1 January 1998; accepted 1 March 1998

Abstract

Recently, techniques motivated by human auditory perception are being applied in main-stream speech technology and there seems to be renewed interest in implementing more knowledge of human speech communication into a design of a speech recognizer. The paper discusses the author's experience with applying auditory knowledge to automatic recognition of speech. It advances the notion that the reason for applying of such a knowledge in speech engineering should be the ability of perception to suppress some parts of the irrelevant information in the speech message and argues against the blind implementation of scattered accidental knowledge which may be irrelevant to a speech recognition task. The following three properties of human speech perception are discussed in some detail:

- limited spectral resolution,
- use of information from about syllable-length segments,
- ability to ignore corrupted or irrelevant components of speech.

It shows by referring to published works that selective use of auditory knowledge, optimized on and in some cases derived from real speech data, can be consistent with current stochastic approaches to ASR and could yield advantages in practical engineering applications.

Zusammenfassung

In jüngster Zeit werden in vielen Bereichen der Sprachverarbeitung Techniken verwendet, die sich an der Verarbeitung im menschlichen Gehör und den Erkenntnissen des menschlichen Sprachverstehens orientieren. Beim Entwurf von Spracherkennungssystemen scheint wieder ein verstärktes Interesse vorhanden zu sein, Wissen über das menschliche Sprachverstehen einfliessen zu lassen. Dieser Artikel gibt die Erfahrungen des Authors bei der Anwendung derartigen Wissens zur automatischen Spracherkennung wieder. Als Grund für die Anwendung solcher Erkenntnisse im Bereich der Sprachverarbeitung ist die Fähigkeit des menschlichen Sprachverstehens zu nennen, einige unwesentliche Informationsanteile in einer sprachlichen Nachricht zu unterdrücken. Es sollten nicht nur vereinzelt weniger wichtige Erkenntnisse verwendet werden, die möglicherweise keine grosse Bedeutung für die Spracherkennung besitzen. Drei Eigenschaften des menschlichen Sprachverstehens werden detailliert erläutert:

- die begrenzte spektrale Auflösung,
- die Verwendung von Informationen über sprachliche Abschnitte, die etwa der Dauer von Silben entsprechen,
- Die Fähigkeit, gestörte oder unwesentliche Merkmale des Sprachsignals nicht auszuwerten.

* Corresponding author. Address: Oregon Graduate Institute of Science and Technology, 20000 NW Walker Road, Beaverton, OR 97006, USA. Tel.: +1 503 690 1136; fax: +1 503 690 1406; e-mail: hynek@ee.ogi.edu

0167-6393/98/$ – see front matter
PII: S0167-6393(98)00027-2

Mit Hinweis auf bereits veröffentlichte Arbeiten wird aufgezeigt, daß die selektive Verwendung des Wissens über das menschliche Sprachverstehen in Einklang steht mit den derzeitigen stochastischen Ansätzen zur automatischen Sprach erkennung und dass dies von Vorteil sein kann in praktischen Anwendungen. Dieses Wissen wird abgeleitet von realen Sprachdaten und mit Hilfe dieser Daten zur weiteren Optimierung verwendet.

Résumé

Récemment, des techniques motivées par la perception auditive, sont appliquées dans de principales technologies courantes de la parole. Il semble y avoir un regain d'intérêt à l'exploitation de plus de connaissance du processus de la parole humaine dans la conception de systèmes de reconnaissance de la parole. Le papier discute l'expérience de l'auteur dans l'application de connaissances auditives à la reconnaissance automatique de la parole. Il avance l'idé que la raison d'appliquer des connaissances de la perception auditive humaine à l'ingénierie de la parole devrait être la capacité de la perception à supprimer quelques parties de l'information contenue dans le message de la parole. L'article plaide contre l'exploitation aveugle de connaissance accidentelle dispersée qui peut être non pertinente pour une tâche de reconnaissance de la parole. Trois propriétés de perception humaine de la parole sont discutées:

- reśolution spectrale limiteé,
- utilisation de l'information contenue dans des segments de longueur d'une syllabe environ,
- possibilité d'ignorer les composantes altérées ou non pertinentes de la parole.

L'auteur montre, en se référant à certains travaux publiés, que l'utilisation sélective de la connaissance auditive optimisée en fonction et dans certains cas provenant de vraies donneés de parole, peut être compatible avec les approches stochastiques actuelles de la reconnaissance automatique de la parole et pourrait avoir des avantages pour des applications pratiques d'ingénierie.

Keywords: Auditory modeling; Human-like processing; Modulation frequency; Automatic speech recognition

1. Introduction

1.1. Knowledge-based ASR – Again?

Human speech communication is a highly specialized task constrained by specific organs involved in the process. Speech production and perception has been and is being studied. However, not much of the acquired knowledge is seen in the design of current automatic speech recognizers (ASRs).

There is no doubt that ASR technology needs further improvement and some believe that the improvement could come from using more speech-specific knowledge in the design of ASR. Whether this is true is a topic of discussion.

Most of successful stochastic ASRs derive their capabilities from extensive training data. Relatively little permanent knowledge is built into the untrained recognizer. Any new application domain requires new training data.

To know more may mean having to learn less. The knowledge built into a design of a recognizer is the knowledge which does not have to be reacquired from the data every time the recognizer is used for a new task.

1.1.1. Some history

Early attempts for large vocabulary continuous speech recognition emerged from noble beliefs that a variability observed in distributions of parameters of phonetic classes could be dealt with by complete understanding of the sources of extra-linguistic variability. Intuitively determined features [1] were derived from the signal and elaborate systems were used to combine information for the final classification. These (sometimes called "knowledge-based") techniques were typically hand-crafted on a particular problem and were mostly applicable only in extremely controlled environments.

[1] Throughout this paper we use the term "feature" to any parameter derived from the speech waveform.

An important paradigm shift occurred in the early 1980s when statistical "ignorance-based" [2] Hidden Markov Model (HMM) based ASR started to dominate the field. A typical HMM ASR system is powerless without extensive training. However, when exposed to a sufficient amount of training data, it typically outperforms knowledge-based ASR approaches.

The dominance of statistical approaches is partly due to the fact that the knowledge of the human speech communication processes is at best incomplete and sometimes outright wrong. Therefore, starting with only a very weakly structured system and letting the training data to provide the rest yields better results than attempts to hardwire the prior knowledge into the design.

Hence, the statistical approaches are not ignorant at all. They recognize the inherent uncertainty of classification, and attempt to derive class decision boundaries from training data. When designed and trained well, they contain true knowledge acquired from real speech data without (possibly wrong) preconceptions about the speech communication process.

When statistical approaches to ASR were introduced, it appeared that the road to progress was to get hands on as much speech data as possible and to be able to process the training set in a reasonable amount of time. Only recently, the speech community is being challenged by the availability of large amounts of "found speech" data [3] and a few problems with current statistical approaches are emerging.

- A classifier trained on large-variance data is only *globally* optimal and does not necessarily perform too well on any of the particular subproblems. [4]
- To optimize the performance, the underlying models are relatively weak (multi-Gaussian). Systems grow large and demands on training data are extensive. Generalization on new unseen problems is quite poor.
- The way the knowledge is represented in the system is not always transparent. Techniques for re-use of the acquired knowledge are not well developed.
- In striving for the best performance, hand-crafting of the model structure and empirical "fudging" of probabilities from sub-systems is common.

The underlying original philosophy of statistical ASR is that "the data should speak". However, since there is currently no principled way to derive the structure of the model from the data, the hand-crafting of the HMM model is in some way reminiscent of procedures from knowledge-based ASR, and makes it vulnerable to similar problems.

It appears that not much of reusable and widely available knowledge comes out from the current extensive ASR research. More time should be devoted to discussing and interpreting the obtained results, reporting efforts which failed to perform to expectations and discuss reasons for such failures. A premium in rewarding ASR research should be paid on efforts and achievements which aim for advancing general knowledge applicable to future ASR problems.

1.1.2. A light at the end of the tunnel

At the moment it seems that at least some sources of undesirable variability can be identified and their effects alleviated. In particular, the effects of variable communication environments are better understood. In some cases, the effect of the undesirable variability in a given feature space is trivial and relatively simple techniques are efficient in reducing it. This leads to works in robust feature extraction, and feature and model adaptation which, by using the knowledge of the problem, to some extent alleviate needs for extensive training on different environments.

However, not all sources of undesirable variability are simple to model and easy to identify and stochastic classification techniques present powerful means for dealing with such sources of random variability. As such stochastic approaches need to be respected, studied, and utilized. However, they should not entirely substitute for better understanding and better utilization of speech-specific knowledge. Today's explosive developments of

[2] The origin of this term is claimed by at least two speech researchers.

[3] Speech recorded from radio and TV broadcasts.

[4] One could in principle compete in both down-hill and cross-country using the same back-country telemark skis, but would it be an optimal strategy?

computing and data acquisition technology should be used for acquiring permanent and reusable knowledge of human speech communication process which could be used in improving practical ASR.

1.2. Where to start?

There is a number of suspicious points in a structure of a typical main-stream large-vocabulary continuous-speech ASR which are inconsistent with the knowledge about speech process.

This paper focuses only on one source of knowledge which is being frequently neglected in ASR and which imposes powerful constraints on human speech communication process, the knowledge of human speech perception.

It is unlikely that significant amounts of linguistic information could be found in speech components which are not heard by an average listener. Thus, better knowledge of the process of human speech perception may point to natural constraints for sensible speech processing. The fact that we communicate by voice suggests the speech signal interpreted by a system of human speech perception may carry enough information for a reasonable self-normalization and practical phonetic classification in the presence of nonlinguistic variability.

1.2.1. Why human speech perception?

The human speech communication process, illustrated in Fig. 1, involves a source (organs of speech production), a channel (environment), and a receiver (organs of speech perception). Communication theory teaches that for optimal communication, all three components should be in tune with each other. It is quite likely that similarly, forces of nature found a way to optimally allocate available resources for human speech communication through an imperfect acoustic communication channel. The historically younger speech signal (generated by organs of speech production which today serve multiple purposes and evolved from organs which originally provided only more basic life-supporting functions) might have evolved to accommodate properties of human auditory perception.

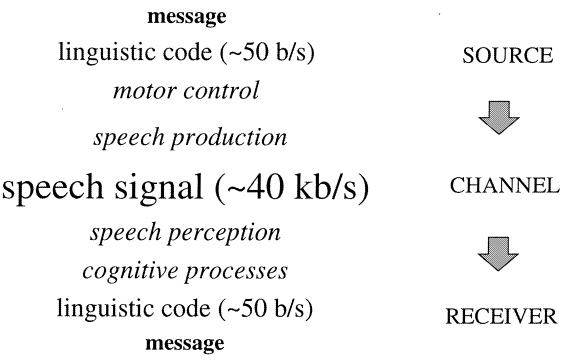

Fig. 1. The information rate in human speech communication process is highest on a speech signal level. An important role of speech perception is to reduce this rate by alleviating some of nonlinguistic variability.

2. Analysis in ASR

2.1. The task

As illustrated in Fig. 1, the information rate in the speech signal is by some estimates [24] of the order of 40 kbits/s. The written equivalent of the linguistic message in the signal is less that 60 bits/s [24]. The general task of ASR is to identify the linguistic information in the signal in presence of other nonlinguistic variability. That is, in effect, to reduce the amount of information by about 3 orders of magnitude! The analysis for ASR should support this goal.

One of the most significant sources of nonlinguistic variability in speech are speaker-dependent factors such as a length of the vocal tract, fundamental frequency of the voice, or acquired habits of a given speaker. Another source is the environment in which the speech is produced, transmitted, and received – the communication channel. The channel introduces additive noise, as well as linear and nonlinear distortions. Ideally, the analysis in ASR should deliver features consistent with the underlying linguistic message without regard for such nonlinguistic factors.

2.2. A typical approach

ASR typically uses features based on a short-term spectrum of speech. Such a representation

describes the time-varying speech signal by a sequence of short-term feature vectors. Each vector reflects properties of a relatively short (10–20 ms) segment of the signal. Each individual feature vector is usually treated as independent of its neighbors.

A majority of stochastic ASR systems attempt to capture temporal aspects of the signal by concatenating individual context dependent phoneme models, each model typically having three different states, representing piecewise approximation of the internal phoneme dynamics. Within each state, the parameters are assumed to be stationary, identically distributed, and independent of the neighboring analysis frames. This assumption implies no constraint on a particular temporal order of speech parameters within each state. Such a piecewise stationary model is a rather crude approximation of the rich temporal dynamics of speech.

2.3. Some of the problems of short-term analysis

We suggest that the concept of short-term analysis as a feature extraction technique for ASR should be revisited.

Firstly, there is some evidence that attention to spectral detail required for high-quality coding of speech may be excessive and not required for ASR applications.

Further, there is a recent notion (discussed later in this article) that robust feature extraction, in order to extract reliable information for classification of sub-word units of speech, needs access to at least about syllable-length (around 200 ms) spans of the speech signal. This may allow for distinguishing between stationary and speech-like nonstationary events in the speech signal and to alleviate some coarticulation effects.

Finally, global features such as cepstral coefficients of spectral envelopes are easily affected by common frequency-localized random perturbations which have hardly any effect on human speech communication. It may be that local descriptors of spectral properties of speech (see e.g. [73]) should be used in conjunction with missing data techniques to avoid catastrophic failures in realistic communication environments.

The short-term speech representation is historically inherited from speech coding applications. Needs in ASR and in speech coding are clearly different. Generally speaking, in speech coding, speaker-dependent and environmental information need to be preserved to permit high-quality resynthesis of speech. In ASR, there is no need to reconstruct the original speech, and most often, the goal is to extract the linguistic message, and alleviate other nonlinguistic factors present in the signal. Therefore, speech analysis techniques for ASR could be [5] quite different from the ones used in speech coding.

Overall, we agree with Allen [2] in speculating that the "across-frequency" processing is one of the causes of the extreme fragility of current ASR in realistic situations and that (consistently with the properties of human speech perception) more attention needs to be paid to the temporal structure of the speech signal in the analysis of speech.

3. Auditory-like analysis in ASR

3.1. Some reasoning for auditory-like analysis

ASR attempts to access the human speech communication process (through a microphone) and to decode the message which was originally intended for the human listener. Processes of human and automatic recognition of speech are illustrated in Fig. 2. Both strive for the same goal – to get the linguistic message from the signal. Both appear to use the similar strategy – to eliminate the nonlinguistic variability in the signal in order to derive the message.

If speech developed so that it would optimally use properties of human auditory perception, it could make some sense that the machine feature extraction should attempt to closely emulate the part of the human communication chain for which the signal is intended, i.e. human speech perception.

[5] But essentially are not.

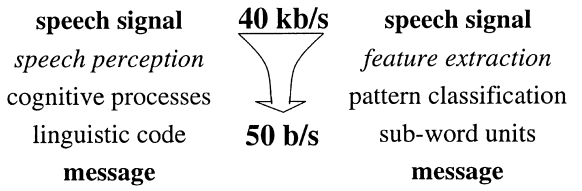

Fig. 2. Role of speech analysis in ASR is similar to a role of speech perception in human communication.

3.2. Why do recognizers not yet have ears?

Thus far, however, auditory models have not yet found full acceptance in ASR. We speculate that there are several possible reasons for this seeming failure. Some of these reasons are listed below.

- Not respecting nature of current ASR which requires features with certain properties.
  - Current ASR prefers features which are independent of each other, identically distributed within a given recognition unit (usually phoneme), with statistics easily described by gaussian mixture distributions. Thus, the auditory-like features may not be suitable for a conventional recognizer or the rest of the recognizer may need to be re-optimized for the new feature representation.
  - New techniques are often tested on a well established task in a system which is finely tuned to some other technique. In a complex ASR system, there are many things that can go wrong, and usually at least one of them does when new technique is substituted for an old one.
- Testing on tasks that do not expose weaknesses of conventional feature extraction techniques.
  - Recognizers work well on clean well controlled laboratory data. Applying auditory models to the tasks where conventional techniques work well may not reveal advantages of auditory-like approaches. Improvements should be sought and expected in real environments, where conventional ASR techniques fail.
- Modeling properties of hearing which may have only a minor role in human speech communication.
  - Not everything what human auditory perception can do is necessarily used in decoding of speech. Most of existing auditory knowledge was derived from experiments with artificial stimuli in well controlled environments. We need more and better knowledge from real speech stimuli at realistic signal and noise levels.
- Last but not least, one of the main reasons for the limited success of auditory modeling in ASR may be a failure to understand the purpose of the front-end processing in ASR. As discussed earlier in this paper, the very purpose of phonetic classification in ASR is a significant reduction of the information carried by the speech signal. Thus, the front end processing should be supportive of this task. This last point is discussed throughout the rest of this paper.

3.3. Ask not what hearing can do for you...

We do not act on all information which is available around us. What might be important for ASR is not so much what human speech perception can get but rather what it does not get or does not use from the acoustic signal. Following such logic, the following two principles were found beneficial for ASR.

- Human hearing has its limits, and due to such limits, certain sounds are perceptually less prominent than others. One reasonable principle is to eliminate what human listeners cannot hear while focusing attention on parts of the signal that are heard well. Later in this paper we discuss two auditory constraints:
  - Limited spectral resolution of the analysis.
  - Temporal masking (dependency of perception of the current sound on preceding sounds).
- The second reasonable principle is to alleviate parts of the signal which, even when heard well, do not supply reliable cues for decoding the message. Some parts of the signal may be corrupted by noise, certain parts of frequency spectrum of the signal may not bear information about identity of a given acoustic segment, etc.

Humans may be able to make decisions about the reliability of a particular part of the signal based on semantics and syntax of the linguistic message.

We currently do not have many tools for determining the reliability of the data but efforts in this direction are ongoing. Later in this paper we discuss an emerging multi-band technique which we believe to be a step towards emphasizing reliable parts of the available data.

4. Spectral domain

Many well-accepted signal processing algorithms used in ASR emulate some of constraining properties of human hearing such as the nonlinear (Bark, Mel) frequency scales [11,48,61,18,34], spectral amplitude compression [56,37,32,39,34], decreasing sensitivity of hearing at lower frequencies (equal-loudness curves) [48,34] and large spectral integration [53,23,22,17] by principal component analysis [65], by cepstral truncation [11,61], or by low order autoregressive modeling [34]. These algorithms, as well as their relevant properties, are discussed in some more detail in [42]. Even the most ardent opponents of auditory-like approaches in ASR use such techniques in their everyday work.

4.1. Perceptual linear prediction (PLP)

4.1.1. The principle

PLP [34] combined several engineering approximations to selected characteristics of human hearing:

1. critical band (Bark) nonlinear frequency resolution, emulated by integrating the short-term Fourier spectrum of speech under increasingly wider trapezoidal curves (sometimes substituted my Mel-spaced triangular filters [75]),
2. asymmetries of auditory filters, emulated by a relatively steep (25 dB/Bark) slope of the trapezoidal curve towards higher frequencies and a more gradual (10 dB/Bark) slope towards lower ones,
3. unequal sensitivity of human hearing at different frequencies, emulated by a fixed approximated Fletcher–Munson equal loudness curve,
4. intensity–loudness nonlinear relation, emulated by a cubic root compression, and

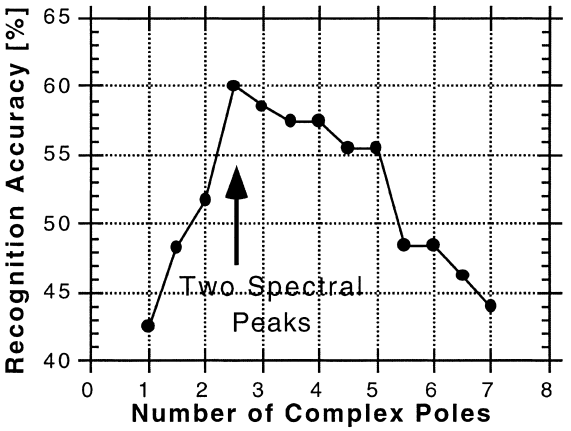

Fig. 3. Accuracy of cross-speaker recognition as a function of number of complex poles of PLP model. When forced to generalize from speech of a single speaker, a two-peak spectral representation yields the best results.

5. broader than critical-band integration, hypothesized in perception of speech (see e.g. [53]), emulated by an autoregressive all-pole model.

All these steps contribute to effectiveness of PLP analysis, the most important being the nonlinear warping of the frequency axis.

4.1.2. The effect of the model order

The effect of model order of the PLP all-pole model was studied experimentally on cross-speaker speech recognition experiments in which training data from one speaker were used to recognize speech of another speaker. As indicated in Fig. 3, the 5th order model was found to be optimal.

Fig. 4 shows that in comparison to the conventional formant based representations, the broader spectral integration done by low-order PLP analysis is capable of delivering more similar results from human and mynah-bird speech. [6] Similarity of PLP results for adult and child speech has been also reported [34].

[6] Evaluating speech analysis on mynah-bird speech may seem far stretched but argument can be made that the analysis for ASR should perform reasonably consistently on all kinds of speech-like sounds acceptable to human listeners.

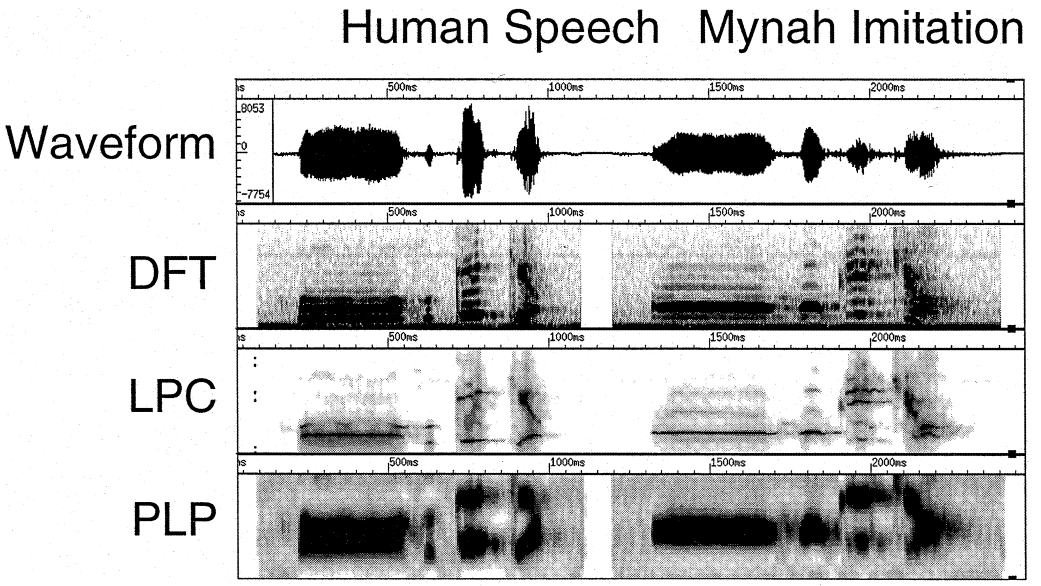

Fig. 4. Spectrum of the same sentence produced by human and by mynah-bird obtained by conventional spectral analyses and by low-order PLP analysis. Peaks in LPC spectra are in quite different positions. Low-order PLP approximates energy clusters in the short-term spectrum and yields more similar representations.

4.1.3. PLP in ASR

The 5th order PLP model was used successfully in speaker-independent recognition of digits [34]. Recently, PLP analysis was reported to be the most efficient speech representation in extensive DARPA evaluations of the large vocabulary continuous speech ASR technology [16,75]. For such complex task with a sufficient amount of training data, a higher model order (DARPA systems typically use model order 12) is more efficient.

4.2. Gross spectral features as carriers of linguistic information

Results of cross-speaker ASR experiments show that the spectrum which contains less detail performs better on this task than a representation with higher spectral resolution (see Fig. 3 adapted from [34]). Such a result suggests that a speaker-dependent information (which is harmful in the cross-speaker ASR) is minimized while a linguistic information is still sufficiently preserved in the gross spectral shape of the speech spectrum. This is further supported by Malayath et al. [59] who reported optimality of low-order spectral representation in representing a linguistic message using stochastic Oriented Principal Component technique on time-aligned multi-speaker data.

Such behavior is consistent with smoothing regularization in stochastic training (see e.g. [28]) and it would be satisfying to discover that the forces of nature found a rigorously justifiable way to deal with the generalization problem in human speech communication. To support this notion, we discuss below some historical evidence which suggests that it is a gross spectral feature structure rather than the fine details of the spectral envelope which is a consistent indicator of the linguistic message across different speakers.

- Ladefoged [55] mentions that young Isaac Newton observed that as a tall glass is filled with liquid (Newton used beer), one hears series of vowels /u/, /o/, /a/, /e/, /i/ (i.e. ordered by a resonance frequency of an uncoupled front cavity of the vocal tract in the production of these vowels).
- Helmholtz [33] studied vowels by striking tuning forks with different pitches in front of a mouth shaped for a proper articulation of a given vowel. His general conclusion was that back vowels

could be simulated by a single resonance, while front vowels are better simulated by two resonances.

- Experiments with a pattern playback vocoder [15] support the notion of two spectral energy peaks as the prime carriers of the linguistic message.
- Fant and Risberg [22] observed that all Swedish vowels can be simulated by two spectral peak synthetic stimuli, provided that their second spectral peak F2' is in a particular position, which does not necessarily coincide with any of the formants.
- Fant's F2' concept is supported by Chistovich [17] who observed that human speech perception appears to integrate spectral components within 3–4 Bark spectral bands, therefore being capable of merging several speech formants.
- Itahashi and Yokoyama [48] found that a second spectral peak of the three-peak (6th order) Mel-LPC model approximates well the perceptual effective second formant F2' (Bladon and Fant, 1978).
- Similarly to the Mel-LPC, the 5th order PLP analysis of 18 synthetic cardinal vowels yields results which agree well with Bladon's and Fant's perceptual experiments [34]. Moreover, the bandwidths of the PLP model preserve information about spread of the underlying formant clusters, thus alleviating a fundamental objection [26,7] to the F2' concept (see [34] for evidence and discussion).
- Broad and Hermansky [40,12] speculate that one of reasons for intelligibility of child speech might be ability of human speech perception to simplify particular formant structure and to focus on global and less speaker-dependent spectral properties which carry the information about the shape of a front part of the vocal tract. They demonstrate a strong correlation ($r = 0.9$) 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 vowels, used in articulatory synthesis of the vowel-like sounds. They also show its weak correlation ($r = -0.18$) with the tract length which indicates a relative speaker-independence of the second spectral peak from the low-order PLP analysis. Later [12] they also show a high correlation of the PLP-estimated F2' with the front cavity resonance estimated from the X-ray micro-beam data.

We may wish to revisit the notion of formants of speech as the most important carriers of the linguistic message. It appears that human speech perception may not be resolving higher formants and that it rather focuses on positions and shapes of whole formant clusters in order to extract the linguistic message from voices of different speakers.

5. Temporal domain

5.1. Current phoneme-based ASR

Standard Hidden Markov Model (HMM) phone based recognition uses a rather crude temporal model of the speech process – the speech signal is assumed to be a sequence of steady-state segments. The segments represent phonemes of speech and the temporal structure of speech is imposed by the built-in structure of the HMM chains. Data can influence this structure only on the segmental level by providing boundaries of the segments.

Within the segments, speech is assumed to be a sequence of equally spaced independent short-term acoustic vectors derived from a stationary stochastic process. Each short-term vector represents about 10–20 ms of speech.

5.2. Beyond 20 ms

In reality, the short-term acoustic vectors within the phoneme segments are clearly correlated over time. In spite of well-known powerful speech production phenomena of coarticulation, auditory perception phenomena of forward masking, and linguistic concept of syllable, all of which point to temporal dynamics over a time interval of the order of several hundreds of ms, the medium-term temporal properties of speech have not been extensively studied and utilized in speech processing.

Only recently, techniques such as multi-vector input [58] combined either with MLP [8,49,72] or

with linear discriminant analysis (LDA) [46,47,13], dynamic (delta) features [27], RASTA processing [38], short-term cepstral mean removal [66], dynamic cepstrum [1], or probabilistic optimum filtering [62], are emerging as post-processing techniques which operate on sequences of the short-term feature vectors. It appears that such post-processing techniques derive their strength from a "locally global" view of speech in which about a syllable-length segment of the speech signal is considered in deriving features for the subsequent classification (see [35] for more discussion).

It is impossible to observe the difference between temporal properties of speech and noise on the basis of a single 20 ms sample from the signal. To be able to differentiate between the slow and fast rates of spectral change one needs to compare spectral information from a relatively large segment of speech (see Fig. 5). Only then it may become apparent that some spurious spectral components (such as the F2 spectral peak in Fig. 5) could exhibit temporal behavior which would be inconsistent with temporal properties of speech components.

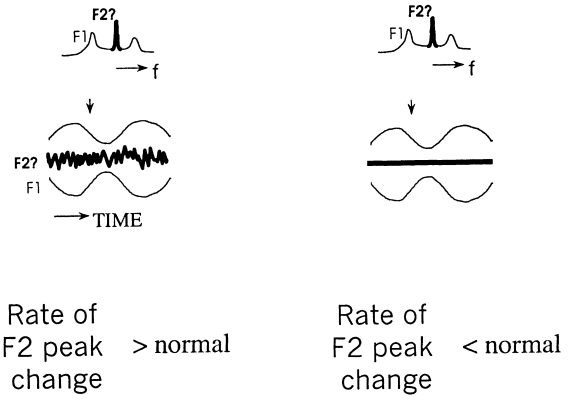

Fig. 5. A single frame from the short-term analysis does not give enough information to determine whether the second spectral peak in the short-term spectrum represents useful spectral component or noise. Only when looking at spectral dynamics on a larger segment of speech it becomes obvious that in the first case the second spectral peak is due to random noise (which characteristics vary faster than speech), in the second case the spectral peak comes from a steady tone which characteristics are constant (or vary slower than speech).

5.3. Dynamic features

The first widely accepted step towards employing temporal structure of speech were so called dynamic cepstral features [27]. The dynamic features can be interpreted as estimates of the first and second temporal derivatives of the time trajectories of cepstral coefficients. They are most often computed as the first and the second orthogonal polynomial expansion of the feature trajectory over some short segment of speech (typically up to 90 ms, although longer segments were also successfully used [27,31]).

Benefits from the use of dynamic features are widely recognized and their success suggests that the dynamics of short-term speech features over extended time intervals provide useful additional information about the linguistic identity of the given time instant in speech.

5.4. RASTA processing of speech

5.4.1. The principle

The RelAtive SpecTrAl (RASTA) technique was originally developed as a purely engineering technique for dealing with fixed or slowly varying nonlinguistic components of speech features. Linear distortions or additive noise in the speech signal may show up as biases in appropriately transformed short-term spectral parameters. Since the rate of such extra-linguistic changes is often outside the typical rate of change of linguistic components, Hermansky et al. and Hirsch et al. [41,44] have proposed that filtering the temporal trajectories of speech parameters might alleviate the extra-linguistic spectral components from the speech representation. This technique came to be known as RASTA speech processing.

The RASTA band-pass filtering is typically done either on the logarithmic spectrum (or cepstrum, which is a linearly transformed logarithmic spectrum) or on the spectrum compressed by $\ln(1 + \text{const} \cdot x)$ nonlinearity.

Recently Hermansky et al. [39] have reported that RASTA filtering on root-compressed power spectrum (with filters designed from the training data) is also effective for perceptual enhancement of noisy telephone speech. Interestingly, the

data-derived filters appear to be enhancing the syllabic rate of speech and supressing components with lower and higher modulation frequencies [6].

5.4.2. The effect of the filter time constant

Series of recognition experiments were run with different RASTA filters to determine the optimal filter structure. Results of one of these experiments are shown in Fig. 6. The optimal filter for recognition of noisy speech is a bandpass filter with the pass-band between about 1 and 12 Hz. The time constant of the integrator in the filter is about 170 ms. RASTA processing enhances dynamic events in the signal and suppresses the steady or slowly varying ones. Fig. 7 illustrates the effect of RASTA processing. Comparing to a conventional short-term analysis such as PLP, RASTA emphasizes changes in the signal.

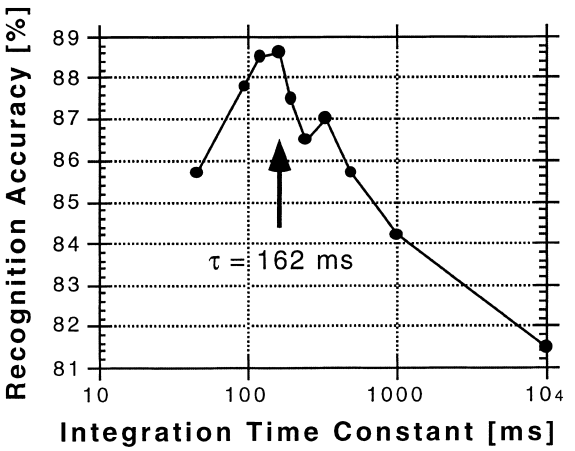

Fig. 6. Dependency of recognition accuracy in presence of linear distortions on a time constant of integrator of RASTA filter. The accuracy is highest when the effective length of impulse response of RASTA filter spans approximately over a syllable.

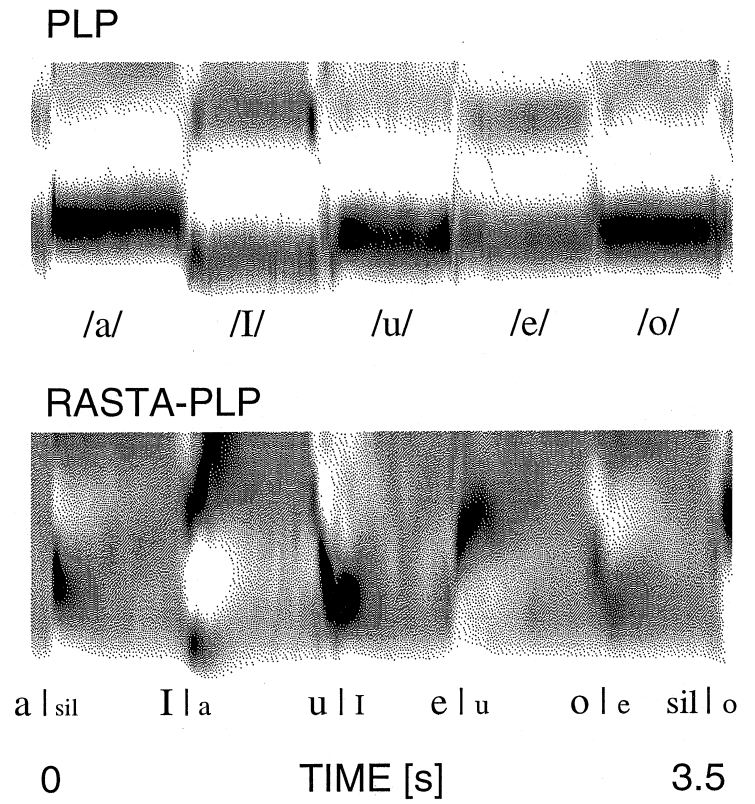

Fig. 7. Spectra of five sustained Czech vowels obtained by PLP and RASTA-PLP analyses. Note enhanced transitions resulting from RASTA processing. The vowels are extra long (about 70 ms each) for the purpose of illustration of the effect.

Even though RASTA is neither the first nor the only auditory model which explicitly emulates temporal properties of human hearing (see e.g. [14,67]), it utilizes larger time-spans of the signal than most of the other models do (impulse response of the whole original RASTA filter [38] has effective length about 220 ms). Use of such rather large time spans points to the whole so far relatively untapped domain of the so-called modulation spectrum of speech and to new possibilities offered there [43,38,35,30]. This is discussed to some extent in the following section.

5.4.3. RASTA in ASR

RASTA works very well with ASRs based on whole-utterance units. With a conventional phoneme-based systems, the benefits are mixed: the system becomes more robust in presence of channel variability but it may slightly deteriorate in good conditions because of the increased dependency on the context of a given speech segment.

In addition, one needs to realize that IIR RASTA filter has in principle an infinite memory (and in practice the filter initializes only after some 250 ms). This causes no principal problems when continuously processing natural speech but can be of concern when processing end-pointed speech without its natural surroundings where the initialization of the RASTA filter is an issue. This may be disturbing to those who are used to a standard short-term analysis where every analysis frame is independent from its neighbors. [7]

However, human auditory perception is not being switched on only when speech comes either and the percept of a current phoneme very much depends on its surroundings too (see e.g. [74] and its companion papers for an excellent evidence). Thus, it is not RASTA which should be blamed. Rather, one should re-design the recognizer to confirm to properties of new feature extraction. [8]

[7] As it was initially disturbing to us too.

[8] That does not imply we currently know exactly how to do that.

5.5. Modulation spectrum of speech

The prime carrier of the linguistic information are changes of the vocal tract shape. Such changes are reflected in changes of the spectral envelopes of the speech signal. The spectral envelope is represented by speech parameters used in ASR (e.g. cepstral coefficients). The patterns of speech parameters vary gradually within each distinct segment of speech and their particular dynamics is an important cue to the identity of a given segment.

Spectral analysis of temporal trajectories of spectral envelopes of speech yields the modulation spectrum of speech [45]. Dominant components of the modulation spectrum indicate the dominant rate of change of the vocal tract shape. The modulation spectrum of speech is dominated by components between 2 and 8 Hz, reflecting the syllabic and phonetic temporal structure of speech [45,30]. The whole concept of the modulation spectrum of speech is illustrated in Fig. 8.

Human auditory system is most sensitive to modulation frequencies around 4 Hz (see e.g. [35] for the review of evidence available in the literature). Thus, human hearing in perception of modulated signals acts as a band-pass filter. The impulse response of such a band-pass filter would need to span at least 150–250 ms. It is tempting to speculate that dominance of components around 4 Hz in the modulation spectrum of speech is a consequence of properties of human hearing.

To obtain sufficient spectral resolution at low modulation frequencies, rather large time spans of speech (of the order of at least several hundreds ms) are required to compute the modulation spectrum of speech.

5.5.1. Perception of speech with modified dynamics of spectral envelopes

Drullman [20,21] shows the dominant importance of the modulation frequencies around 4–6 Hz for speech communication. Inspired by Drullman's experiments, Arai et al. [3] have initiated a series of similar experiments using a slightly different signal processing paradigm which is based on the residual-excited LPC vocoder and aiming for band-pass processing of trajectories of spectral envelopes [3]. Their results confirm and extend

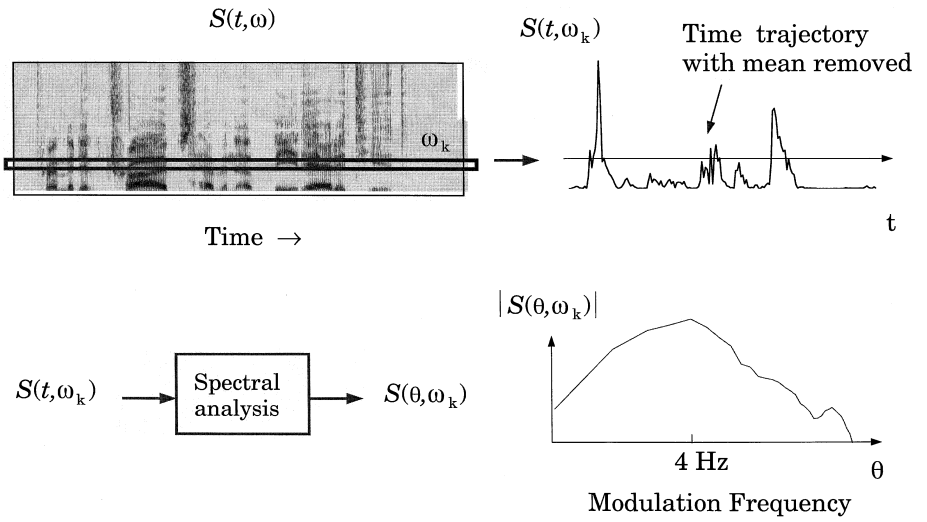

Fig. 8. Modulation spectrum of speech describes spectral components of time trajectory of spectral envelope of speech. The modulation spectrum of continuous uninterrupted speech is dominated by a syllabic rate of speech which is typically close to 4 Hz.

Drullman's. It appears that the modulation frequencies between DC and 1 Hz, as well as those above 15 Hz, play only secondary roles in speech communications.

5.5.2. ASR with modified dynamics of spectral envelopes

Arai's perceptual experiments were a starting point for a similar series of ASR experiments using narrow band-pass filtering of time trajectories of cepstral coefficients of speech [51]. Results of both the perceptual and ASR experiments are summarized in Fig. 9, which is adopted from [52]. This figure shows the relative importance of various components of the modulation spectrum of speech for human and machine communication. Details of the technique used for obtaining this data are given elsewhere [51,52] but the overall conclusion from the experiments is that the 2–8 Hz part of the modulation spectrum is the most important for speech communication while the 0–1 Hz band, as well as components higher than 16 Hz play only a secondary role in human communication and might be harmful for ASR, especially when contaminated by environmental noise.

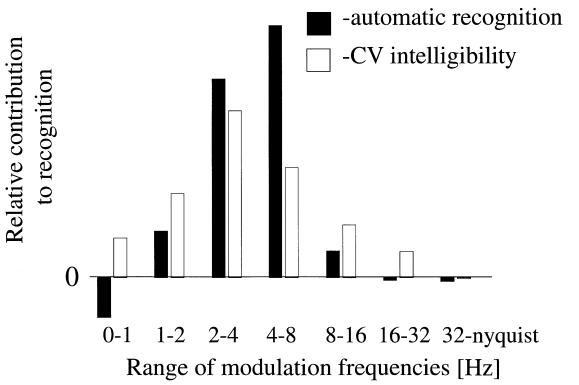

Fig. 9. Relative importance of components of modulation spectrum of speech for human and machine communication. Both intelligibility of speech and performance of ASR are most severely degraded when components in a vicinity of 4 Hz are attenuated. Extremely low (below 1 Hz) and high (above 16 Hz) components of the modulation spectrum have only minor role in human speech communication. Alleviating components below 1 Hz appears to be beneficial in ASR.

5.6. Data-driven design of RASTA filters

The initial ad hoc form of the RASTA filters was optimized on a relatively small series of ASR experiments with noisy telephone digits. The

optimizations using ASR experiments are costly and there is no guarantee that the solutions obtained will not be specific to a given ASR system. Therefore, we designed a data-based optimization which avoids using a specific ASR paradigm. Such a technique is based on the linear discriminant analysis (LDA) which is a stochastic technique, that attempts to optimize the linear discriminability between classes in the presence of undesirable within-class variability (see e.g. [46,13] for some examples of previous use of LDA in ASR). There is a way to structure the LDA problem in such a way that the LDA solution can be interpreted as a set of FIR RASTA-like filters which are applied on time trajectories of spectral energies. The process of deriving FIR RASTA filters by LDA is illustrated in Fig. 10.

The input vector was formed from segments of a time trajectory of a single logarithmic critical-band energy over a relatively long (about 1 s) span of time. Each vector typically spans more than a single phoneme, and is labeled by the phoneme at the center of the vector. Having formed such 101-dimensional (each vector spans about 1 s at 100 Hz sampling) vector space with vectors labeled by their respective phoneme classes, LDA analysis yields a 101×101 scatter matrix, decomposed into its principal components. Then the principal vectors represent FIR filters which most efficiently (with respect to the within-class and the across-class variability) map the 101-dimensional input space to several points of the output space.

The first discriminant vector, the most important for the discrimination, typically explains about 80–90% of the variability in the data. The second discriminant vector then explains anywhere between 10–15% of the variability, and the third discriminant vector still explains more than 5% of the variability. Thus, the properties of those higher discriminant vectors are also of interest.

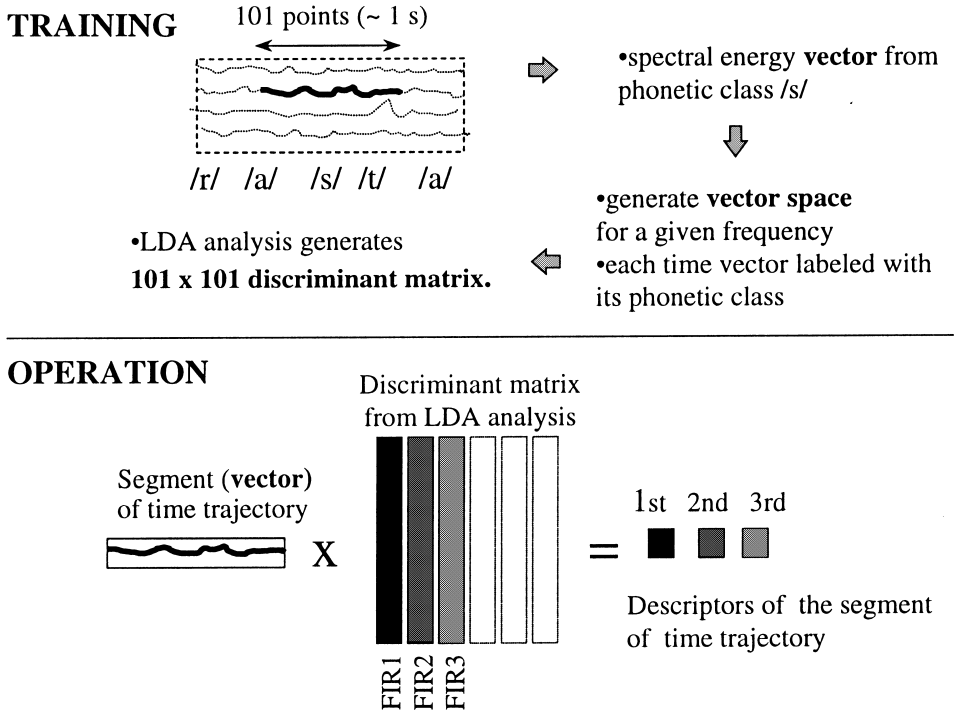

Fig. 10. Data-driven design of FIR RASTA filters using Linear Discriminant Analysis. In such a setup, each column vector of the LDA-derived discriminant matrix is a RASTA-like FIR filter applied to a time trajectory of a short-term critical-band spectral energy.

In initial experiments we have used about 30 min of hand-labeled conversational telephone speech (switchboard database). To allow the data-driven RASTA filter to compensate for harmful source of variability, the fixed convolutional distortion (which shows as an multiplicative constant in the spectrum of speech) was simulated in the training data by multiplying each spectral energy value by a constant, emulating the 9 dB linear distortion. This artificially distorted data were added to the original database. So, the database contained both the undistorted and the distorted speech.

Frequency characteristics of the first three discriminant filters from this design are shown in the upper part of Fig. 11 by full lines. Frequency characteristics of the original RASTA filter, and of the third and second orthogonal polynomials approximating the time trajectory of the feature within a nine frame (90 ms) time interval as proposed by Furui [27] are shown in the bottom part of the figure by dotted lines. The impulse responses are shown in Fig. 12. Only the filters operating on the critical band with a center frequency of 5 Bark are shown here but the filters for other critical bands are very similar. The peak at 1 Hz in the frequency response of the second discriminant filter appears to be a consequence of a slight DC bias in its impulse response.

So far we have been using as classes context-independent phonemes from the three-hour OGI Stories hand-labeled database, 20 min of the hand-labeled Switchboard database, and 6 h of the forcefully aligned Switchboard database. The general characteristics of the data-derived RASTA filters are relatively independent of the particular database used for their design.

It appears that for the optimal linear discrimination of speech into context-independent phonetic classes one needs to use features derived from at least 250 ms long segments of the signal. The most important discriminant feature is derived by weighting such speech segment by a function which can be reasonably well emulated by a difference of two gaussians [60] with σ_1 about 100 ms and σ_2 about 20 ms. Such function implies a mild

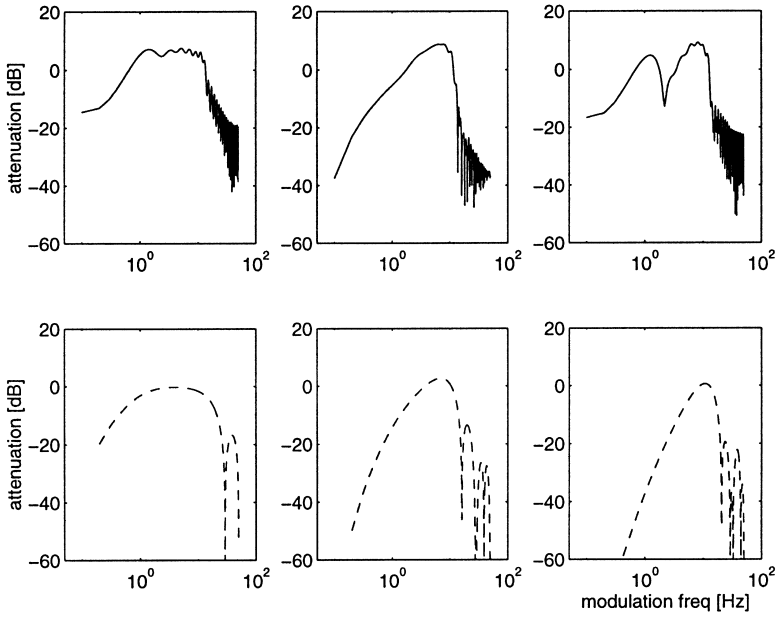

Fig. 11. Frequency characteristics of first three discriminant vectors from LDA design using about 30 min of phoneme-labeled Switchboard data with simulated additional linear distortions (full lines at the top) and of the original RASTA filter, the second orthogonal polynomial (slope) over 90 ms of time trajectory combined with the RASTA filtering, and the third orthogonal polynomial over 90 ms (curvature) combined with the RASTA filtering (dotted lines at the bottom).

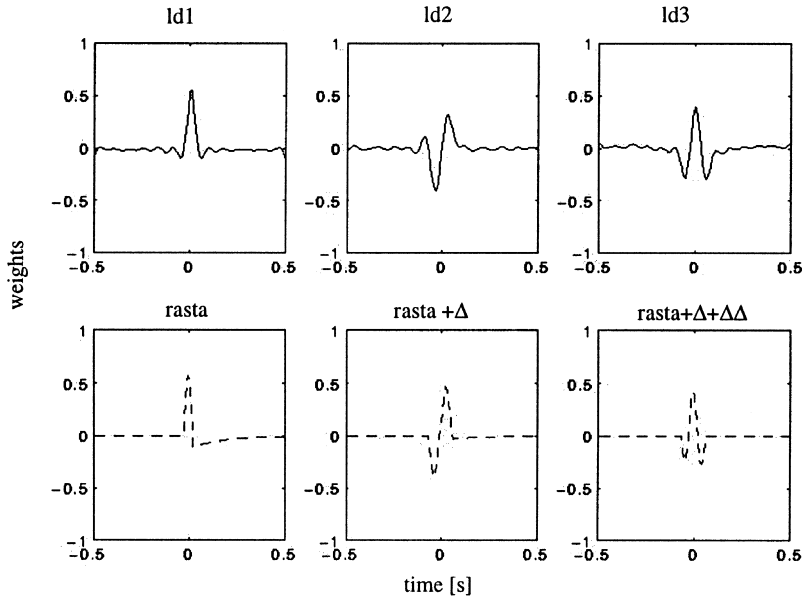

Fig. 12. Impulse responses of the first three vectors from LDA and of the original RASTA filter, the second orthogonal polynomial (slope) over 90 ms of time trajectory combined with the RASTA filtering, and the third orthogonal polynomial over 90 ms (curvature) combined with the RASTA filtering (dotted lines at the bottom).

temporal lateral inhibition. The second most important weighting function is reasonably well approximated as a time-derivative of the first one and the third one can be obtained by a double temporal differentiation of the impulse response of the first filter. All three weighting functions suppress steady and slowly varying components, as well as fast changing components of the time trajectories of spectral energies. Such functions are consistent with recent findings in neurophysiology of auditory cortex [73,4,19].

Quite different constrained optimization technique was also applied with similar aims and similar results to the design of RASTA filters from data [5].

The data-driven RASTA filters yield an advantage in ASR over the conventional RASTA filter [71].

5.7. "Sluggishness" could be good

The argument is that speech parameters should reflect instantaneous configurations of the vocal tract. Therefore, temporal resolution of a conventional short-term speech analysis is typically of the order of tenths of ms so that the frame-by-frame analysis can closely follow changes in the vocal tract shape. We are suggesting above that speech features should be derived from time intervals of at least an order of magnitude larger. Such a notion deserves a short discussion.

The instantaneous properties of the speech signal may be important in current techniques for speech synthesis and coding. However, what counts from the communication point of view is the effect of the acoustic event on the receiver, i.e. on the human speech perception system (or the ASR).

The effect of even short acoustic event on human auditory perception is likely to last for a rather substantial amount of time. It appears that the short-term memory of the auditory periphery in mammals (exhibited by forward masking (see, e.g., [76]), or the buildup of loudness (see, e.g., [68])) is of the order of about 200 ms. Recent physiological measurements from the auditory cortex of mammals are consistent with these psychophysical observations [73,4,19]. This means

that the human auditory system could in principle utilize rather large (about syllable sized) time spans of the audio signal, just as do the speech analysis techniques discussed here.

The access to and utilization of about a syllable-length segments of speech does not necessarily mean that the analysis would lose an ability to act upon short events in the signal. It merely means that features derived by such analysis at any given time instant would reflect information from a larger segment of speech and that even very short event in the signal would show in the speech representation for an extended length of time. This is illustrated in Fig. 13 which shows spectrograms of a word containing the unvoiced stop /k/ from both conventional PLP analysis (which is a conventional short-term analysis technique with temporal resolution of about 10 ms) and RASTA-PLP (with its temporal resolution of the order of about 200 ms). In RASTA-based "sluggish" analysis, short acoustic events such as stop-bursts spread their influence over a longer time, thus possibly aiding in the identification of such short acoustic events.

Allowing for access to larger segments of speech during analysis could be also advocated using speech production-based arguments [54,69] since it may allow for use of mechanical constraints and principles guiding the underlying speech production process.

5.8. Forward masking and RASTA processing

This section shows consistency of RASTA processing with some auditory effects of forward masking.

If a loud sound is followed closely in time by weaker sound, the perceptability of the weaker sound is diminished. This effect, called forward masking, seems to last (independently of the masker amplitude) for about 200 ms (see e.g. [76]). Forward masking effect is typically measured by presenting a masker (tone or band-passed noise for

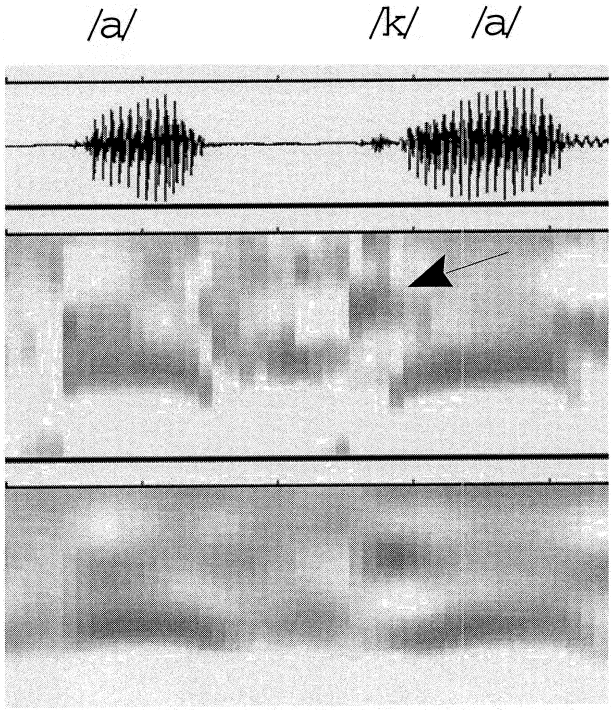

Fig. 13. Spectrograms of a word containing unvoiced stop /k/ from the conventional PLP analysis (upper spectrogram) and the RASTA-PLP (lower spectrogram) with implied time constant about 220 ms. In the "sluggish" RASTA-PLP, the short acoustic events spread their influence into a subsequent speech segment.

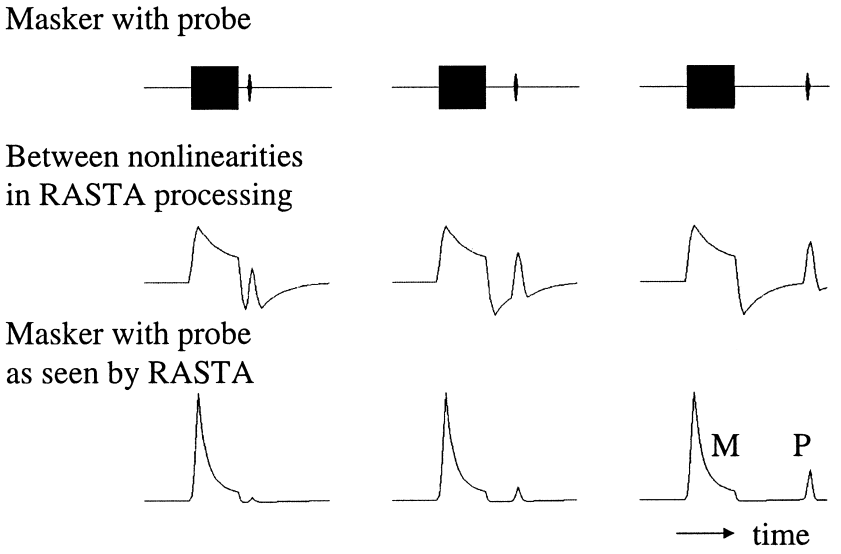

Fig. 14. Mechanism of forward masking in RASTA processing. RASTA filtering is done between a compressive and expansive nonlinearity. RASTA processed masker brings the level of probe down and effectively attenuates it (see also Fig. 15). This effect is greater when the probe closely follows the masker.

200 ms or longer), followed by a short signal called probe. Human observers are asked to detect brief probe presented after a variable delay following the offset of the masker. The masking effect is measured by the rise of a threshold of detection of the probe. The decaying dependence of the masking effect on the log delay is well approximated by a set of straight lines that intersect at a point corresponding to the delay of approximately 200 ms [50]. (See the left part of Fig. 16.)

A model of the dependence of the masking effect on the delay and the masker level (a bilinear relation) requires an essential nonlinearity. Prior attempts to account for the data led researchers to models based on automatic gain control such as proposed by Pavel [63]. In his model, the effect of the masker was to reduce temporarily the system gain. A decade later, Pavel and Hermansky [64] demonstrated that RASTA processing can emulate human perceptual data from forward masking paradigms.

Fig. 14 may help in explaining the process of masking in RASTA processing. Since the length of a masker in a typical forward masking experiment (see e.g. [50]) is comparable to a time constant of the RASTA filter, the level of the RASTA-processed masker at the end of the masker is significantly diminished. Turning off the masker causes a negative shift in the signal. Then, a probe coming shortly after the end of the masker is on lower level than it would be if there was no masker. Since the whole process is sandwiched between compressive and expansive nonlinearities, the negative shift of the compressed probe results in its effective attenuation, as illustrated in Fig. 15.

To evaluate the process quantitatively, we carried out the following experiment. The probe detection was mediated by a comparison of an Euclidean distance between two spectral representations: one representing a loudness [9] temporal profile of the masker alone and another representing temporal profile of loudness of the masker followed by a probe. If there was no probe, the distance between these two loudness profiles of two identical maskers would be zero. Because of the probe in one of the signals, the distance is nonzero. The smaller the processed probe, the

[9] Loudness vectors were computed as a short-term critical-band power spectral energies compressed by the cubic-root nonlinearity.

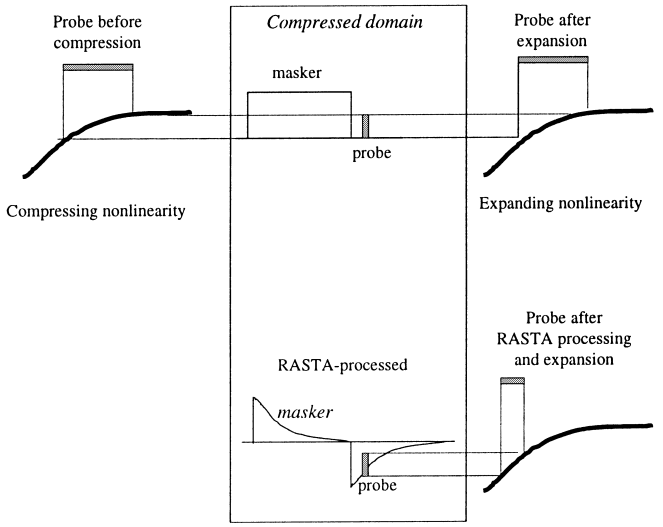

Fig. 15. Attenuation of probe by shift of its level between compressive and expansive nonlinearities.

smaller the distance. Thus, the inverse of the distance is proportional to the amount of masking of the probe.

Results, shown in the right part of Fig. 16, are qualitatively consistent with conclusions from human forward masking experiments [50] which implications are indicated in the left part of the figure.

6. Recognition based on partial information

6.1. A conventional ASR

The first step in most of the current ASR is to convert the incoming speech signal into a series of short-term vectors. Each element of the vector

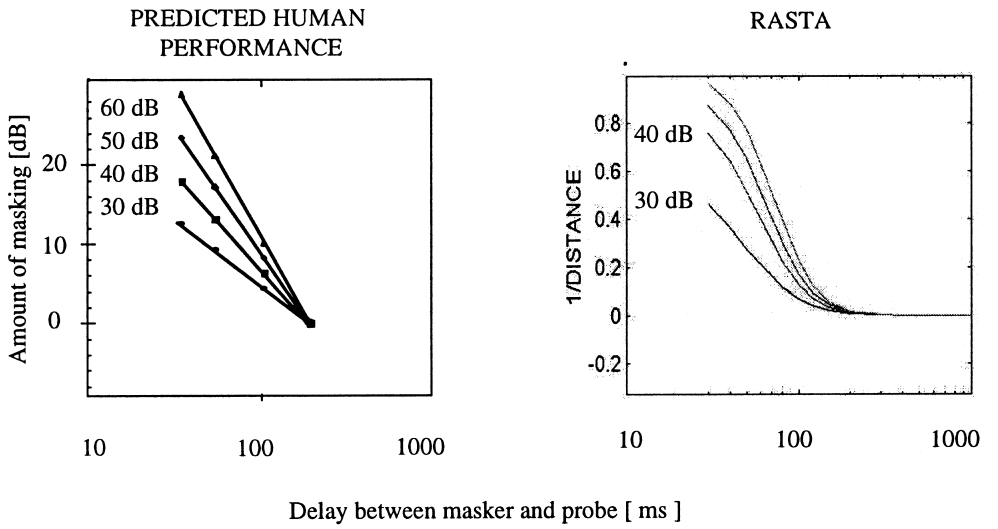

Fig. 16. Extrapolated human data on forward masking (left) and results of experiment with emulation of forward masking using RASTA processing (right).

describes some part of the information carried by the signal; for example each element of the short-term spectral vector represents energy of the speech signal in a given frequency range. Suppose that some of the elements of the short-term vector contain corrupted or misleading information, while the remaining ones are still uncorrupted. This can occur e.g. when the speech signal gets corrupted by selective noise. In the current mainstream ASR the entire feature vector is used as one entity and even a single corrupted spectral element can severely degrade the performance of the recognizer.

6.2. The human way

Fletcher's early work [25] on the Articulatory Index (see [2] for a review) suggests that human auditory perception works differently from current ASR. Specifically, Fletcher suggests that the linguistic message gets decoded independently in different frequency sub-bands and the final decoding decision is based on merging the information from the sub-bands. According to Fletcher, the probabilities of erroneous recognition in the sub-bands $P\{E_i\}$ multiply to yield the overall error rate $P\{\text{Error}\} = \prod_i P\{E_i\}$. One interpretation of Fletcher's work is that as soon as any sub-band combination yields sufficient information, the information from the remaining (possibly corrupted) sub-bands does not have to be used for subsequent decoding of the message.

The fact that the spectral envelope can be so easily corrupted by common distortions such as linear filtering or additive environmental noise, which have only a minimal adverse effect on human speech communication, may put in a question the whole concept of spectral analysis for deriving an internal representation of acoustic signal in human cognition. Even though there is a strong evidence that human auditory perception does some sort of spectral analysis of the incoming acoustic signal, it may be that the main reason for frequency selectivity of human auditory system is not to derive frequency content of a given segment for phonetic classification but rather to provide means for optimal choice of high signal-to-noise (SNR) regions for deriving reliable sub-band based features by temporal analysis of the high SNR sub-bands of the signal. Such view would be supported by some recent findings in physiology of auditory cortex [73].

6.3. Multi-band ASR

Hermansky et al. [36] and Bourlard et al. [9,10] examine Fletcher's proposal by subdividing the available speech spectrum into a number of frequency sub-bands and extracting spectral features from each of the sub-bands. Recognition is done independently in each of the sub-bands; each recognizer yielding conditional probability estimates for all the classes to be recognized (see Fig. 17). These estimates are then merged to give the final result. In our work, merging is done by a multilayer perceptron (MLP) trained on the training data.

Experiments with frequency selective noise (reported in [36]) show the potential advantage of the multi-band approach in ASR of noisy speech. An important observation from these early experiments, summarized in Fig. 18, was that when leaving parts of the speech spectrum out of the ASR process, the ASR accuracy degrades relatively slowly. This may be suggesting that linguistic information in the sub-bands is to some extent redundant and, consistently with [57], access to just a few sub-bands may facilitate reasonable speech communication. Such an observation is consistent with [29] and is a good news for robustness in ASR because it means one can leave out corrupted parts of the speech spectrum without severely impairing the ASR process.

Hermansky and Tibrewala [36] investigated several techniques for determination of corrupted sub-bands in the multi-band ASR. Sometimes it is feasible to determine a frequency-dependent signal-to-noise ratio directly from the signal and then leave out the sub-bands in which the level of noise is too high. Otherwise, attempts can be made to determine reliability of data from outputs of the sub-band classifiers. Good estimates of a posterior probabilities should be low when the data are unreliable. Bourlard et al. [10] thus used a simple linear combination of sub-band classifier outputs to perform the merging and reported good results

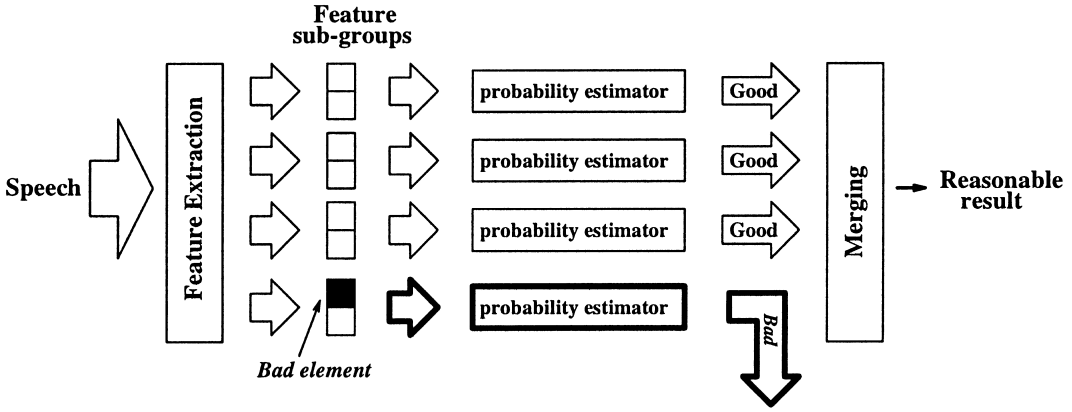

Fig. 17. Principle of multi-stream ASR. A potential advantage of this approach is in situations when the individual sub-streams carry different information about the underlying communication process and are differently affected by disturbances.

in ASR of noisy speech without any need for the prior determination which sub-bands are noisy. Subsequently, Tibrewala and Hermansky [70] used a nonlinear classifier trained on clean speech data to deal with the corrupted sub-bands. They demonstrated a significant improvement in performance (of about 50% in average) in presence of frequency selective additive noises which corrupted only some of the frequency sub-bands. The multi-band was found ineffective only for noises which corrupted all sub-bands.

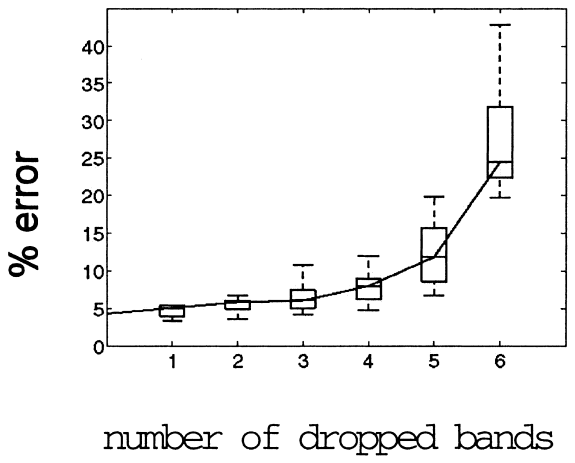

Fig. 18. Accuracy of ASR using partial spectral information. Performance of ASR degrades relatively slowly even when only a small part of the available spectrum is used in ASR. The result shows that e.g. when only 30% of the available spectrum is used (2 sub-bands out of 7) the minimal error increase is only from 4% to 7%.

7. Discussion and conclusion

It is perception which ultimately determines which components of the signal are used for a decoding of the message. Thus, any gross inconsistencies between properties of ASR technology and properties of human hearing should be viewed with caution.

Emulation of certain properties of human hearing in feature extraction for ASR can be useful. Some of the main-stream techniques such as Mel or Bark auditory-like frequency warping of speech spectrum are used virtually by everybody. Some of the emerging ones were discussed in this paper.

This paper tried to convey the idea that indiscriminate use of accidental knowledge about human hearing in ASR may not be what is needed. Not all properties of human hearing are relevant to speech communication, and some are not even well understood. Lessons from the past teach that using the wrong knowledge can be worse that using no knowledge at all. So the question is how to select and how to use the relevant knowledge.

Straight emulation of well established properties of human hearing into engineering systems is one way to provide a priori knowledge into the ASR problem. Reasonable processing constraints

may be introduced this way. Current dominant speech analysis techniques in ASR such as Mel cepstral analysis or PLP make good use of the prior knowledge about the nonlinear frequency scale, spectral amplitude compression, and the equal loudness curve.

Some other powerful properties of human auditory perception, such as temporal masking, or the ability to efficiently use corrupted information, are just starting to be exploited, and we have discussed initial attempts for their use in this paper.

Proponents of purely engineering approaches to ASR may find that a proper use of auditory constraints may be justified on engineering grounds. Thus, e.g. reducing the number of free parameters in low-order PLP is consistent with regularization theory with limited amount of training data, the use of multi-stream ASR could be argued from the point of view of missing features theory in robust statistical processing, etc.

One should not underestimate the power of real speech data in deriving and implementing auditory knowledge. Speech developed to optimally use properties of human auditory perception and relevant auditory properties are reflected in the structure of the acoustic speech signal. Thus, the data should not be used as a lazy-man's substitute for knowledge but rather as a potential source of permanent and reusable knowledge.

A conventional way to use the data is to adjust the parameters of the model so that the performance of the system on development data improves. Such a technique yielded low-resolution PLP modeling for deriving speaker independent features, or the RASTA band-pass filter which reduces effects of noise.

However, an other effective way is to avoid the ASR altogether and apply optimization techniques on a simplified task, relevant to the ASR goal. This has been done by optimizing phoneme-class separability in our LDA-based RASTA filter design.

It is gratifying to start to believe that there may not be a dilemma between data-driven and knowledge-based approaches after all: the data carry the relevant knowledge. It has been imprinted on the data by forces of nature in attempts to optimize the human speech communication process.

It is hard to accept that speech technology could not benefit from better understanding of speech communication process. This understanding can come by studying either part of the process: the speech production system, the signal, and the human speech perception. They are all closely related and properties of one part are reflected in properties of an another. This paper, while focusing on human speech perception, makes an ample use of the signal (e.g. in the data-driven design of RASTA filters), as well as looking for a possible support in speech production (e.g. in the relation between F2' and resonance frequency of the front part of the vocal tract [40,12]).

While this paper mainly discussed possible modifications of the analysis process, it is possible that even larger gains can be found in modifying (or even abandonment) the currently dominant but perceptually inconsistent pattern-matching approach to ASR. Such paradigm attempts to assign each segment of the signal to a class taken from a close set of classes. Human communication is not a uniform process in which all parts of the signal are carrying an equal amount of information about the message. Rather, instants of high information content are interleaved with large portions of the signal carrying practically no useful information at all. Attempts for recognizing everything in the signal may thus be wasteful or even counterproductive.

But that would be yet another story...

7.1. Should airplanes flap wings?

It is said that car does not need to have legs to move and airplanes do not need to flap wings to fly, *ergo* ASR does not need to have ears to recognize speech. We in principle support such view, believing that progress should be made by the *knowledge of the principle guiding a process* rather than by *copying the appearance of the process*.

Today's ASR resembles early designs of machines heavier-than-air: just as those early machines, it appears clumsy and is seems to consist of randomly collected pieces of machinery – no matter how hard we try, the technology is still not up to the task.

Some attempts for auditory modeling in ASR may resemble the legendary Icarus with bird-like flapping wings – blindly applying observations from nature but quite obviously not understanding the basic laws which make the process work.

Only when the Wright brothers and their contemporaries came to understand that it is not flapping the wings but rather the Bernoulli force which keeps birds in the air, modern aviation was born. We as speech technologists should also strive for understanding which properties of human auditory perception are relevant for decoding the speech signal and are likely to improve performance of ASR in realistic situations.

Acknowledgements

This work has benefited from collaboration and discussions with many friends and colleagues. The interactions with Morgan, Steven Greenberg, and Herve Bourlard of ICSI, David Broad, Gentleman Scientist in Santa Barbara, Jordan Cohen of IDA Princeton, Misha Pavel of OGI, Frederic Jelinek of Johns Hopkins University, and B. Yegnanarayana of IIT Madras, provided many insights presented in this paper. The multi-band approach was inspired by discussions with Jont Allen during the 1993 DoD Workshop on Frontiers in Speech Processing. Mynah bird speech was obtained from Tohru Ifukube of Hokkaido University. Thanks to Nagendra Kumar of Johns Hopkins University for making his LDA program available and for adopting it to our needs. My students and postdoctoral fellows Takayuki Arai, Carlos Avendano, Noboru Kanedera, Narendranath Malayath, Sangita Tibrewala-Sharma, and Sarel van Vuuren did most of the real work behind results mentioned in this paper. Comments of reviewers considerably helped to improve the paper. Finally, I thank the organizations that have provided support for our work over the years, especially the DoD under MDA 904-94-C-6196, the National Science Foundation and ARPA, through IRA-9314959, and Texas Instruments and Intel through their support of our Anthropic Signal Processing Laboratory at OGI.

References

[1] Aikawa, K., Singer, H., Kawahara, H., Tohkura, Y., 1993. A dynamic cepstrum incorporating time-frequency masking and its application to continuous speech recognition. In: Proceedings of the International Conference on Acoust. Speech and Signal Processing, Minneapolis, MN, pp. II-668–671.

[2] Allen, J.B., 1994. How do humans process and recognize speech?. IEEE Trans. Speech Audio Process. 2 (4), 567–577.

[3] Arai, T.M., Pavel, H.H., Avendano, C., 1996. Intelligibility of speech with filtered time trajectories of spectral envelopes. In: Proceedings of the International Conference on Spoken Language Processing, Philadelphia, pp. 2490–2493.

[4] Attias, H., Schreiner, C.E., 1998. Coding of naturalistic stimuli by auditory midbrain neurons. In: Advances in Neural Information Processing Systems, Vol. 10. Morgan Kaufmann, Los Altos, CA.

[5] Avendano, C. van Vuuren, S., Hermansky, H., 1996. Data based filter design for RASTA-like channel normalization in ASR. In: Proceedings of the International Conference on Spoken Language Processing, Philadelphia.

[6] Avendano, C., Hermansky, H., 1997. On the properties of temporal processing for speech in adverse environments. In: Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk Mountain House, New Paltz, New York.

[7] Bladon, A., 1983. Two-formant models of vowel perception: Shortcomings and enhancements. Speech Communication 2, 305–313.

[8] Bourlard, H., Morgan, N., 1994. Connectionist Speech Recognition – A Hybrid Approach. Kluwer Academic Publishers, Dordrecht.

[9] Bourlard, H., Hermansky, H., Morgan, N., 1996. Copernicus and ASR challenge: Waiting for Kepler. In: Proceedings of the ARPA ASR Workshop Spring 1996, Arden House, NY, pp. 157–162.

[10] Bourlard, H., Dupont, S., 1996. A new ASR approach based on independent processing and re-combination of partial frequency bands. In: Proceedings of the International Conference on Spoken Language Processing, Philadelphia, pp. 426–429.

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

[12] Broad, D., Hermansky, H., 1989. The front cavity/F2' hypothesis tested by data on tongue movements. J. Acoust. Soc. Amer. 86 (Suppl. 1), S13–S14.

[13] Brown, P., 1987. The acoustic-modeling problem in automatic speech recognition. Ph.D. Thesis, Computer Science Department, Carnegie Mellon University.

[14] Cohen, J.R., 1989. Application of an auditory model to speech recognition. J. Acoust. Soc. Amer. 85 (6), 2623–2629.

[15] Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M., Gerstman, L.J., 1952. Some experiments on the perception

of synthetic speech sounds. J. Acoust. Soc. Amer. 24, 579–606.
[16] Cook, G.D., Christie, J.D., Clarkson, P.R., Hochberg, M.M., Logan, B.T., Robinson, A.J., 1996. Real-time recognition of broadcast radio speech. In: Proceedings of the International Conference on Acoust. Speech and Signal Processing, pp. 141–144.
[17] Chistovich, L.A., 1985. Central auditory processing of peripheral vowel spectra. J. Acoust. Soc. Amer. 77, 789–805.
[18] Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28 (4), 357–366.
[19] deCharms, C.R., Blake, D., Merzenich, M.M., 1997. Sound feature decomposition by the primary auditory cortex. In: 1997 Workshop on Advances in Neural Information Processing, Breckenridge, Colorado (submitted to Science, also unpublished technical memo).
[20] Drullman, R., Festen, J.M., Plomp, R.,1994. Effect of temporal envelope smearing on speech reception. J. Acoust. Soc. Amer. 95, 1053–1064.
[21] Drullman, R., Festen, J.M., Plomp, R., 1994. Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Amer. 95, 2670–2680.
[22] Fant, G., Risberg, A., 1962. Auditory matching of vowels with two formant synthetic sounds. Quarterly Progress and Status Report 4, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm.
[23] Fant, G., 1965. Acoustic description and classification of phonetic units. Ericsson Technics, No. 1, reprinted in: Fant, G., 1973. Speech Sounds and Features. MIT Press, Cambridge, MA.
[24] Flanagan, J.L., 1972. Speech Analysis Synthesis and Perception, second edition. Springer, Berlin.
[25] Fletcher, H., 1953. Speech and Hearing in Communication. Krieger, New York.
[26] Fujimura, O., 1964. On the second spectral peak of front vowels: A perceptual study of the role of the second and third formants. Language and Speech 10, 181–193.
[27] Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal Process. 29, 254–272.
[28] Geman, S., Bienenstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilemma. Neural Computation 4 (1), 1–58.
[29] Green, P.D., Cooke, M.P., Crawford, M.D., 1995. Auditory scene analysis and hidden Markov model recognition of speech in noise. In: Proceedings of International Conference on Acoust. Speech and Signal Processing, Detroit, MI, pp. 401–404.
[30] Greenberg, S., 1997. On the origins of speech intelligibility in the real world. In: Proceedings of ESCA-NATO Tutorial and Research Workshop on Robust speech recognition for unknown communication channels, Pont-a-Mousson, France, pp. 23–32.
[31] Hanson, B.A., Applebaum, T.H., Junqua, J.C., 1996. Spectral dynamics for speech recognition under adverse conditions. In: Lee, C.H., Soong, F.K., Paliwal, K.K. (Eds.), Automatic Speech and Speaker Recognition. Kluwer Academic Publishers, Dordrecht.
[32] Hanson, B., Wong, D., 1984. The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence in interfering speech. In: Proceedings of International Conference on Acoust. Speech and Signal Processing, pp. 18.A.5.1–18.A.5.4.
[33] Helmholtz, H., 1954. On the Sensation of Tone. Dover, New York.
[34] Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Amer. 87 (4), 1738–1752.
[35] Hermansky, H., 1995. Exploring temporal domain for robustness in speech recognition. In: Proceedings of the 15th International Congress on Acoustics, Vol. II, Trondheim, Norway, pp. 61–64.
[36] Hermansky, H., Tibrewala, S., Pavel, M., 1996. Towards ASR on partially corrupted speech. In: Proceedings of International Conference on Spoken Language Processing, Philadelphia, PA, pp. 462–465.
[37] Hermansky, H., Fujisaki, H., Sato, Y., 1983. Analysis and synthesis of speech based on spectral transform linear predictive method. In: Proceedings of International Conference on Acoust. Speech and Signal Processing, Boston, MA, pp. 777–780.
[38] Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Process. 2 (4), 578–589.
[39] Hermansky, H., Wan, E., Avendano, C., 1995. Speech enhancement based on temporal processing. In: Proceedings of International Conference on Acoust. Speech and Signal Processing, Detroit, MI, pp. 405–408.
[40] Hermansky, H., Broad, D., 1989. The effective second formant F2' and the vocal tract front cavity. In: Proceedings of International Conference on Acoust. Speech and Signal Processing, Glasgow, Scotland, pp. 480–483.
[41] Hermansky, H., Morgan, N., Bayya, A., Kohn, P., 1991. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). In: Proceedings of Eurospeech'91, Genova, Italy, pp. 1367–1371.
[42] Hermansky, H., Pavel, M., 1995. Psychophysics of speech engineering systems. Invited paper, 13th International Congress on Phonetic Sciences, Stockholm, Sweden, pp. 42–49.
[43] Hermansky, H., 1988. Modulation spectrum in speech processing. In: Prochazka, A., Uhlir, J., Rayner, P.J.W., Kingsbury, N.G. (Eds.), Signal Analysis and Prediction. Birkhauser, Boston.
[44] Hirsch, H.G., Meyer, P., Ruehl, H., 1991. Improved speech recognition using high-pass filtering of subband envelopes. In: Proceedings of Eurospeech'91, Genova, Italy, pp. 413–416.
[45] Houtgast, T., Steeneken, H.J.M., 1985. A review of the MTF concept in room acoustics and its use for estimating

speech intelligibility in auditoria. J. Acoust. Soc. Amer. 77 (3), 1069–1077.
[46] Hunt, M.J., 1979. A statistical approach to metrics for word and syllable recognition. J. Acoust. Soc. Amer. 66 (S1), S35(A).
[47] Hunt, M., Lefebvre, C., 1989. A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In: Proceedings of International Conference on Acoust. Speech and Signal Processing, Glasgow, Scotland, pp. 262–265.
[48] Itahashi, S., Yokoyama, S., 1976. Automatic formant extraction utilizing mel scale and equal loudness contour. In: Proceedings of International Conference on Acoust. Speech and Signal Processing, Philadelphia, PA, pp. 310–313.
[49] Janseen, R.D.T., Fanty, M., Cole, R.A., 1991. Speaker independent phonetic classification in continuous English letters. In: Proceedings of International Joint Conference on Neural Networks, Seattle, WA, pp. II-801–808.
[50] Jestead, W., Bacon, S.P., Lehman, J.R., 1982. Forward masking as a function of frequency, masker level, and signal delay. J. Acoust. Soc. Amer. 950–962.
[51] Kanedera, N., Arai, T., Hermansky, H., Pavel, M., 1997. On the importance of various modulation frequencies for speech recognition. In: Proceedings of Eurospeech'97, Rhodos, Greece, pp. 1079–1082.
[52] Kanedera, N., Arai, T., Hermansky, H., Pavel, M., 1997. On the relative importance of various components of the modulation spectrum for automatic speech recognition. Submitted to Speech Communication.
[53] Klatt, D.H., 1982. Speech processing strategies based on auditory models. In: Carlson, R., Granstrom, B. (Eds.), The Representation of Speech in The Peripheral Auditory System. Elsevier Biomedical Press, New York, pp. 181–202.
[54] Kozhevnikov, V.A., Chistovich, L.A., 1967. Speech: Articulation and Perception. Translated from Russian by US Department of Commerce, p. 250, 251.
[55] Ladefoged, P., 1967. Three Areas of Experimental Phonetics. Oxford Univ. Press, Oxford, p. 65.
[56] Lim, J.S., 1979. Spectral root homomorphic deconvolution system. IEEE Trans. Acoust. Speech Signal Process. 27 (3), 223–233.
[57] Lippmann, R.P., 1995. Accurate consonant perception without mid-frequency speech energy. IEEE Trans. Speech and Audio 4 (1), 66–69.
[58] Makino, S., Kawabata, T., Kido, K., 1983. Recognition of consonant based on the perceptron model. In: Proceedings of International Conference on Acoust. Speech and Signal Processing, Boston, MA, pp. 738–741.
[59] Malayath, N., Hermansky, H., Kain, A., 1997. Towards decomposing the sources of variability in speech. In: Proceedings of Eurospeech'97, Rhodos, Greece.
[60] Marr, D., 1982. Vision. Freeman, San Francisco, CA.
[61] Mermelstein, P., 1976. Distance measures for speech recognition, psychological and instrumental. In: Chen, R.C.H. (Ed.), Pattern Recognition and Artificial Intelligence. Academic Press, New York, pp. 374–388.
[62] Neumayer, L., Weintraub, M., 1994. Probabilistic optimum filtering for robust speech recognition. In: Proceedings of International Conference on Acoust. Speech and Signal Processing, Adelaide, Australia, pp. I-417–420.
[63] Pavel, M., 1980. Homogeneity in complete and partial masking. Ph.D. Thesis, New York University.
[64] Pavel, M., Hermansky, H., 1994. Temporal masking in automatic speech recognition. J. Acoust. Soc. Amer. A 95, 2876.
[65] Pols, L.C.W., 1971. Real-time recognition of spoken words. IEEE Trans. Comput. 20 (C) 972–978.
[66] Rosenberg, A.E., Lee, C., Soong, F.K., 1994. Cepstral channel normalization techniques for HMM-based speaker verification. In: Proceedings of International Conference on Spoken Language Processing, Yokohama, Japan, pp. 1835–1838.
[67] Seneff, S., 1985. A joint synchrony/mean-rate model of auditory speech processing. J. Phonetics 16 (1), 55–76.
[68] Stevens, J.C., Hall, J.W., 1966. Brightness and loudness as functions of stimulus duration. Perception and Psychophysics 1, 319–327.
[69] Stevens, K.N., 1996. Applying phonetic knowledge to lexical access. In: Proceedings of Eurospeech'95, Madrid, Spain, p. 3.
[70] Tibrewala, S., Hermansky, H., 1997. Multi-band and adaptation approaches to robust speech recognition. In: Proceedings of Eurospeech'97, Rhodos, Greece, pp. 2619–2622.
[71] van Vuuren, S., Hermansky, H., 1997. Data-driven design of RASTA-like filters. In: Proceedings of Eurospeech'97, Rhodos, Greece, pp. 409–412.
[72] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K., 1988. Phoneme recognition using time-delay neural networks, Proceedings of International Conference on Acoust. Speech and Signal Processing, New York, pp. 107–110.
[73] Wang, K., Shamma, S.S., 1995. Spectral shape analysis in the central auditory system. IEEE Trans. Speech Audio Process. 3 (5), 382–394.
[74] Watkins, A.J., Makin, S.J., 1997. Some effects of filtered context on the perception of vowels and fricatives. J. Acoust. Soc. Amer. 99 (1), 588–594.
[75] Woodland, P.C., Gales, M.J.F., Pye, D., 1996. Improving environmental robustness in large vocabulary speech recognition. In: Proceedings of International Conference on Acoust. Speech and Signal Processing, pp. 65–68.
[76] Zwicker, E., 1975. Scaling. In: Keidel O., Neff W. (Eds.), Handbook of Sensory Physiology, Vol. V.3. Springer, Berlin, pp. 401–448.