SPOONS '90:
The SPeech recOgnition frOnt eNd workShop

N. Morgan[1], H. Hermansky[2]
and C. Wooters[1]

TR-90-045

September 11, 1990

Abstract

An appropriate input representation is crucial for pattern classification. In spite of this, we find that feature extraction, transformation, and selection tend to be under-represented aspects of the speech recognition literature. Therefore, the authors decided to gather together a group of interested parties for a dialog on the subject. We ultimately invited a group of about 30 researchers, and on July 6, 1990, held a 1-day workshop which we called SPOONS. This document is a brief summary of that day, including the abstract for each talk.

[1]International Computer Science Institute, 1947 Center Street, Suite 600 Berkeley, CA 94704-1105, USA
[2]U S WEST Advanced Technologies, Englewood, CO

Session I: Features and Transformations

The basic theme of this session was speech representations which were found to be discriminant, or otherwise useful for later processing. A particular target was the naive view that feature extraction is unnecessary for neural net classification; "give the net a speech waveform and it will learn everything else." Agreement was easily reached that such an approach is hopeless, and there was anecdotal information about some failed attempts at this approach. However, a proposed experiment on the ICSI RAP machine would test out completely automatic feature learning on a large speech database. This is planned for the Fall.

The speakers for this section were Nelson Morgan, who discussed feature selection for pattern recognition; Ron Cole, who explored the issues of speech signal representations in general (e.g., what is important, what is not, how can we compare them, and should we set up some way to talk further); and Les Atlas, who discussed his approach to "Truly Nonstationary Time-Frequency Analysis of Speech" [Atlas et al, 1990].

Some of the better (approximate) quotes from this session:

Morgan: Throwing away information is often better than keeping it.

Cole: The first person you need to convince in scientific investigation is yourself.

Cole: We need an analysis technique which is fast even without a Cray mainframe and which gives visible results.

O'Malley: Why a ten millisecond analysis step? Because we have ten fingers?

Cole: We have public-domain speech evaluation databases; how about having a public-domain speech feature extraction software? We already have cochleagrams and Perceptual Linear Prediction (PLP) in the Oregon Graduate Institute (OGI) package and we are happy to share it with anybody.

Atlas: Our analysis [Generalized Time-frequency Representation] gives good spectral and temporal details; one can even see the instant of glottal opening. Could that be useful in speech recognition?

At one point during this discussion, John Ohala mentioned that the "impulse" which drives the resonant vocal tract for voiced speech is the glottal closing, rather than the glottal opening. Somebody in the audience requested explanation, since the air from the lungs is the energy source. John Ohala gave a rather involved explanation, followed by this clarification from Mike O'Malley:

O'Malley: Which is louder? This (clapping hands together), or this (pulling them apart)?

A procedural conclusion of the morning meeting was that we should get together again. An annual meeting was suggested, as well as some sort of electronic mailing list. These ideas are currently being worked on.

Session II: Psychoacoustic Models

Psychoacoustic models try to approximate *properties of hearing*, not necessarily its mechanism. Thus, they may model both the lower-level (peripheral) and higher-level (neural) processing stages of human hearing.

The two models presented at the workshop were the Modulation Spectrum model [Pueschel, In Prep] and Perceptual Linear Prediction (PLP) model [Hermansky, 1990]. The Modulation Spectrum model attempts to explain phenomena of hearing such as temporal masking and segregation of different acoustic sources. Study of its applications in ASR is planned for the near future. The PLP model attempts to extract the linguistic information from speech and was developed specifically for speaker-independent ASR applications. Its use in the ICSI speech recognition system is currently under investigation.

The discussion in this session brought up several interesting issues:

J. Cohen: The value of a new model is often not immediately obvious but could become evident when the model is applied in some new area.

Hermansky: Agreed. First we just tried to substitute our PLP analysis in the conventional LPC - based ASR. The advantage, if any, was only marginal. Later, in conjunction with the group delay metric, the smoothness of PLP-derived spectral envelope turned out to be one of the more important advantages over conventional speech analysis techniques. We think that the absence of drastic frame-to-frame changes in PLP may allow for an efficient use of spectral dynamic features.

Lazzaro: Referring to efforts for linking some perceptually-based features with important gestures of the vocal tract in speech production, it should be possible for a neural-like electronic structure to find an automatic mapping procedure between the acoustic speech signal and muscle activity during speech production, thus bypassing the vocal tract shape completely.

[Someone]: Is speech special? In music we often listen to "personality" more than listening for the "message". Why is it that we could tell the difference between speakers but we tend to ignore it when it comes to decoding the linguistic information?

J. Cohen: We could play speech through a filter approximating the inverse of a steady vowel spectrum (such as 'e') and the speech is still intelligible, including the vowels which turn into a white spectrum signal. Which hearing model can account for that?

This latter comment provoked a good deal of discussion, and is currently driving some of the joint research between the authors (at ICSI and US West); we would like to design a model that could survive Jordan Cohen's "inverse e" test.

Session III: Physiological Models

Physiological models attempt to emulate the functional mechanisms of the human auditory system. (Note: the distinction between physiological models and psychoacoustic models is fuzzy at best.)

There were two models of this type presented. The first was presented by Dick Lyon and Malcom Slaney of Apple and is called a Correlogram [Slaney and Lyon, 1990]. This approach models the physiology of the outer and middle ear up to the inner hair cells. The second physiological model was presented by Shihab Shamma. Shamma's model, which was based on experimental evidence from single unit neural recordings, focuses on the mappings of a signal across the primary auditory cortex [Shamma, 1988].

Here are a few audience comments to give the flavor of the discussion:

General objection to physiological models: Why do we want to model the functional mechanisms of the human auditory system? After all, the way that we do speech recognition may not be the best way for the computer to do it.

Response : Although the mechanisms that we use in hearing may not be the best mechanisms for the computer to use, it is not necessarily a bad idea to model them. Besides, we can learn something about ourselves in the process.

Comment: There are a number of intriguing similarities between the auditory and visual systems, such as the apparent neural computations of simple functions (such as a Mexican hat function).

Discussion

As might be expected from such a meeting, there was a great deal of stimulating discussion, but no real "answers". Participants who were more oriented towards building speech recognition systems tended toward finding low-dimensional representations which could be shown to be useful, regardless of the dissimilarity with models of human biology. Researchers with stronger interests in modeling physiology tended to prefer representations with higher dimension, which might be more like biological systems in some sense, but which were not necessarily more discriminant. All present, however, seemed to agree that current speech recognition systems were in some sense "incorrect", in that they failed for simple cases in which human speech perception did not. Probably the most memorable example of this was Jordan Cohens's "inverse e" example. In a similar vein, it was agreed that analysis techniques were in their current form because of engineering convenience (i.e., "we know how to use this software package"), because of their mathematical tractability (i.e., "finally even I understand it") or for historical reasons (i.e., "why re-analyze the whole database?"), and not for more fundamental reasons.

Another recurrent theme was that of going beyond simple spectral measures. Both from the psychoacoustic perspective (Pueschel's "Modulation Spectrum") and a more physiological one (Lyon & Slaney's Correlogram), it was agreed that there were temporal features which conveyed important information for the human recognition process. Atlas's GTFR also showed the ability to retain fine temporal features. It was a common

view that some form of this temporal information would need to be used in a system that approached human performance.

These latter issues were discussed somewhat further the next day by a few hardy participants who met to discuss possible follow-up experiments. The primary conclusion was to begin incorporating Hermansky's Perceptual Linear Prediction (PLP) analysis [Hermansky, 1990] in a simple recognizer under development at ICSI [Morgan & Bourlard, 1990], and to then consider expansions to temporal features. Additionally, we agreed to try the "waveform only" approach, given the availability of a fast machine. This latter experiment is scheduled for later on in the year.

The Future of SPOONS

As mentioned in the description of the first session, a common sentiment was that this meeting should be repeated. One possibility that arose was to incorporate it as part of an ungoing IEEE conference, such as the Signals, Systems, and Computers annual meeting at Asilomar. This is currently being considered.

References

L. Atlas, R. Cole, W.Koolman, and P.Loughlin, ''New Nonstationary Techniques for the Analysis and Display of Speech Transients'', *Proc. IEEE Intl. Conf. on Acoustics, Speech, & Signal Processing*, pp. 385-388, Albuquerque, New Mexico, 1990.

D. Pueschel, ''The Modulation Spectrum: Principles and Potential Applications'' ICSI Technical Report, In Prep

H. Hermansky, ''Perceptual linear predictive (PLP) analysis of speech'', J. Acoust. Soc. Am. 87 (4), April, 1990

M. Slaney and R. Lyon, ''A Perceptual Pitch detector'' *Proc. IEEE Intl. Conf. on Acoustics, Speech, & Signal Processing*, pp. 357-360, Albuquerque, New Mexico, 1990.

N. Morgan and H. Bourlard, ''Continuous Speech Recognition Using Multilayer Perceptrons with Hidden Markov Models'', *Proc. IEEE Intl. Conf. on Acoustics, Speech, & Signal Processing*, pp. 413-416, Albuquerque, New Mexico, 1990.

S. Shamma, ''The acoustic features of speech phonemes in a model of auditory processing: Vowels and unvoiced fricatives'', J. Phon. 16, pp. 79-91, 1988

Appendix A: SPOONS Abstracts

Feature Selection for Pattern Recognition

Nelson Morgan
International Computer Science Institute
Berkeley, California

Selection of discrimination-relevant features has been shown to be a critical element of pattern recognition systems [Viglione, 1970; Hanson and Applebaum, 1990]. In particular, selecting useful features and discarding poor ones can yield significant improvement in performance. This will be illustrated with examples from earlier systems which did voiced-unvoiced-silence discrimination[Gevins and Morgan, 1984], and brain wave classification [Gevins and Morgan, 1988]. The relationship to newer schemes for parameter dimensionality reduction such as Optimal Brain Damage [Le Cun, 1990] will be discussed.

References

S. Viglione, "Applications of Pattern Recognition Technology,,, in *Adaptive Learning and Pattern Recognition Systems:Theory and Applications,* Mendel & Fue, eds., Academic Press, 1970

B. Hanson, and T. Applebaum, "Robust Speaker-Independent Word Recognition using Static, Dynamic and Acceleration Features: Experiments with Lombard and Noisy Speech", *Proc. IEEE Intl. Conf. on Acoustics, Speech, & Signal Processing,* Albuquerque, New Mexico, 1990.

A. Gevins, and N. Morgan, "Ignorance-based systems", *Proceedings,* IEEE Intl. Conf. Acoustics, Speech & Signal Processing, San Diego, 39A.5.1–39A.5.4., March, 1984

A. Gevins, and N. Morgan, "Applications of neural-network (NN) signal processing in brain research", *IEEE ASSP Trans. Vol. 36 (7), pp.1152-1161, July, 1988*

Y. Le Cun, J. Denker, and S. Solla, "Optimal Brain Damage",in *Advances in Neural Information Processing Systems II,* Morgan Kaufmann, 1990

Issues Involved in Selecting Signal Representations for Speech Recognition

Ron Cole
Oregon Graduate Institute

My goal is to stimulate discussion on issues related to the selection, evaluation and comparison of signal representations for speech recognition. After a brief presentation, designed mainly to portray the complexity of the problem, the following questions that will be raised for discussion:

(1) What are the desired properties of a signal representation?

(2) Is it possible to meaningfully compare representations? If so, what experimental procedures are likely to produce the most meaningful results?

(3) How can we cooperate to make representations, tools for studying representations, and databases for evaluating representations available to the research community?

Truly Nonstationary Time-Frequency Analysis of Speech

Les Atlas
Dept. of Electrical Engineering, FT-10
University of Washington
Seattle, WA 98195

The usual techniques of processing speech use either a parametric (e.g. LPC) or a non-parametric (e.g. spectrogram) frequency analysis technique of representing stationary intervals of the signal. An implicit assumption in this processing is that there is no change in the frequency content of the signal for a duration of from, say, 3 to 20 mSec. One important effect of the assumption is that there is an inherent time-frequency trade-off which, for example, smears short-duration events such as bursts. Less known, yet important, deleterious effects of this assumption occur in voiced speech where glottal waveform effects are significantly smeared and formant frequency resolution is decreased.

Our own research is based on using the outstanding performance of the auditory system to question the limits of the time-frequency trade-offs of standard spectrogram and LPC approaches. We have adapted the truly non-stationary time-frequency representation of Cohen to provide a representation with outstanding time and frequency resolution and minimal spurious interference terms [Zhao et al, 1990]. This time-frequency representation has been shown to precisely mark burst locations [Atlas et al, 1990], and we will also show our more recent results on the precise location of both formants and the time of glottal opening in voiced speech. Other theoretical results will be presented which give the basis for improved frequency resolution and better performance in noise. Our current research is directed toward showing how our representation can actually improve the performance of speech analysis and recognition systems. We are also investigating other useful properties of time-frequency representations for speech analysis and recognition and their possible impact on the design of truly non-stationary processing techniques.

References

L. Atlas, R. Cole, W.Koolman, and P.Loughlin, "New Nonstationary Techniques for the Analysis and Display of Speech Transients", *Proc. IEEE Intl. Conf. on Acoustics, Speech, & Signal Processing*, pp. 385-388, Albuquerque, New Mexico, 1990.

Y. Zhao, L. Atlas, and R. Marks, "The Use of Cone-Shaped Kernels for Generalized Time-Frequency Representations of Nonstationary Signals". *IEEE ASSP Trans. Vol. 38 (7), pp.1084-1091, July, 1990*

Loudness Adaptation and the Modulation Spectrum in Human Hearing

D. Pueschel
U. of Göttingen and ICSI

A hearing model based on psychoacoustics will be proposed as well as its possibilities to use it for speech recognition.

We developed a model to describe the human hearing system based on psychoacoustic data. Our model addresses the following problems: nonlinear transformation of sound with the effect of masking, temporal integration of sound, backward and forward masking, high temporal resolution of modulations in sound, and segregation to different sound sources.

The approach is to simulate the information loss in the human hearing system introduced by the nonlinear transformations. To compare the model with psychoacoustic data, we simulate thresholds in an adaptive procedure, comparable to the measurements in our human subjects. The first stage is a simple one-dimensional basilar membrane simulation with half-wave rectification. Thereafter follows an adaptive stage, which performs a logarithmic transformation in the static case. In principle this is an automatic gain control, which tends to have a more linear transmission for really short time intervals (1 ms). The adaptive behavior is adjusted with two parameters and includes time constants from 5 to 500 ms. Part of the model is also an analysis into different modulation frequencies after adaptation.

This overall model describes simultaneous and nonsimultaneous masking in every band and is compared with detailed psychoacoustic measurements. The modulation analysis can be done by simple band pass filters which select the different modulation components up to the carrier in every band. This analysis mainly just the Fourier Transform of an autocorrelogram. The bandwidth of the modulation filters determines the behavior for temporal integration, which can be simulated.

It is easy to select a sound source in the representation of sound as a modulation spectrum. We need only to select one modulation component and its harmonics in every band (as long as we have a tonal component). Psychoacoustic experiments about sound segregation are compared with model predictions.

We used the principles of this model for pitch detection algorithms and binaural hearing aids. The possibilities to extract only the speech relevant information will be discussed.

In Search of Linguistic Information in Speech

Hynek Hermansky
U S WEST Advanced Technologies
Englewood, Colorado

We know that spectra of phonetically identical speech segments uttered by different speakers can be quite different. Current ASR systems deal with this variability by training on a large sample of the population, and possible overlap between classes, caused by the speaker-dependent variations, is addressed by sophisticated pattern matching techniques during the recognition. The presented talk describes our struggle in search for speaker-independent features in speech.

First, we look into speech production. Studying x-rays of adult and children vocal tracts in speech production we find indications that the front part of the vocal tract is more speaker-invariant across speakers of different ages than the rest of the tract.

Next, we turn our attention to speech perception. In hope to find the trace of the speaker-invariant front cavity in the auditory-like representation of the speech signal, we apply our Perceptual Linear Predictive (PLP) model. (The PLP representation yields a picture of speech different from the traditional formant-based spectrogram. The large-scale spectral integration is inherent to low-order PLP. The higher peak of the PLP model is estimating the perceptual effective second formant F2'.) By analyzing speech produced with known shape of the front part of the vocal tract we observe that: 1) In experiments with front vowels generated by the articulatory synthesizer, frequency of the second peak of the PLP model remains invariant at the resonance frequency of the front cavity, even though the formant patterns of speech change with changes in the overall tract length. 2) In experiments with real speech, frequency of the second PLP peak is used in predicting position of the vocal tract constriction obtained by the x-ray microbeam in production of sonorant sentences.

Finally, we discuss several parametric sets derived from PLP analysis for use in automatic speech recognition and present some recognition results.

Perceptual Representations of Speech--
Cochleagrams versus Correlograms

Richard F. Lyon
Malcolm Slaney
Apple Computer
Cupertino, CA 95014

We discuss the use of the cochleagram and the correlogram in speech and sound recognition. The cochleagram represents a sound as a pattern of neural firing probabilities at places along the Basilar Membrane inside the cochlea, versus time. It is roughly analogous to the spectrogram and its benefits have been described in several papers. But using the cochleagram as a basis for speech recognition is only a weak way to use knowledge of human auditory processing.

A better representation of speech and sound is called the correlogram. Sound is represented as a two dimensional picture versus time--the extra dimension allows several interesting perceptual experiences to be modeled. We have found the correlogram to be a very rich representation and we have been especially struct by the similarity of auditory and visual percepts. This talk will emphasize the advantages of a two dimensional representation of sound and describe several auditory maps that might be used by the brain to do auditory scene analysis.

Representation of Acoustic Features in the Auditory Cortex

Shihab Shamma
U. of Maryland

Experimental and theoretical studies reveal new mappings of the acoustic signal across the primary auditory cortex. Specifically, besides the binaural columns already known, there are two spectral features that are extracted and encoded in the responses of cortical neurons: The first is the local gradient of the spectrum at each frequency, i.e. there is a two dimensional map where frequency is represented on one axis, and gradient value along the other. A second feature - the direction of FM sweeps at each frequency - is also extracted and mapped in a superimposed representation across the primary auditory cortex. These two maps are direct analogs of similar features that are encoded in the visual cortex, namely the orientation and direction of motion maps.

Appendix B: Conference Participants[1]

Victor Abrash (Stanford, SRI)
Les Atlas (U of Washington)
Mike Berkovec (Stanford)
Jordan Cohen (CCR Princeton)
Michael Cohen (SRI)
Ron Cole (Oregon Graduate Institute)
Gabriel Cristobal (ICSI)
Richard Duda (SJSU)
Jerry Feldman (ICSI, UCB)
Ervin Hafter (UCB)
Hynek Hermansky (US West)
Brian Kingsbury (ICSI, UCB)
Phil Kohn (ICSI)
Wolfgang Kuepper (ICSI/Siemens)
John Lazzaro (CalTech)
Yueming Li (BARRA)
Richard Lyon (Apple)
Peter Marvit (Hewlett-Packard, UCB)
Bernard Mont-Reynaud (Stanford)
Nelson Morgan (ICSI)
Hy Murveit (SRI)
John Ohala (UCB)
Manjari Ohala (SJSU)
Mike O'Malley (Berkeley Speech Works)
Steve Omohundro (ICSI)
Bruce Parnas (UCB)
Dirk Pueschel (ICSI and U of Goettingen)
Charles Della Santina (UCB)
Peter Schwerkhardt (BARRA)
Shihab Shamma (U of Maryland)
Elizabeth Shriberg (UCB)
Malcolm Slaney (Apple)
Joyce Tang (UCB)
Mitchell Weintraub (SRI)
David Wessel
George White (Apple)
Chuck Wooters (ICSI, UCB)
Walter Yamada (UCB)
Jun Zhang (ICSI, U of Wisconsin)

---

[1]This is the sign-up list that was passed around at the meeting. It is thus a snapshot of one point in the day; while most participants stayed for the entire day, some others may have missed this list.