

**The Dynamics of Speech Production and Perception (a NATO Advanced Study Institute)**

**Keith Bromley**

**June 24 - July 6, 2002**

---

*These reports summarize global activities of S&T Associate Directors of the Office of Naval Research International Field Offices (ONRIFO). The complete listing of newsletters and reports are available on the ONRIFO homepage: <http://www.onrifo.navy.mil/> from the Newsletter link, under the authors' by-line, or by email to respective authors.*

**Keywords**

*speech, hearing, speech processing, signal processing, human factors*

T A B L E O F C O N T E N T S

**Introduction**

**Technical Focus**

**Defense Relevance**

**Conference Content by Speaker:**

What Are the Essential Clues for Understanding Spoken Language?

Time Frames of Spoken Language

- Dr. Steven Greenberg, University of California at Berkeley, USA

Neural Dynamics of Speech Perception

- Dr. Stephen Grossberg, Boston University, USA

Evidence for Multi-Resolution Analysis of Auditory Stimuli

- Dr. George Meyer, Keele University, United Kingdom

Acoustic/Modulation Frequency Transforms for Single-Channel Talker Separation

- Prof. Les Atlas, University of Washington, USA

The Auditory Image Model and its Use in Speech Analysis

Time-Domain Auditory Processing of the Dynamic Aspects of Speech

- Prof. Roy Patterson, Cambridge University, United Kingdom

**Assessment**

**Contacts**

## Introduction

The NATO Advanced Study Institute on the "Dynamics of Speech Production and Perception" was held on June 24 - July 6, 2002 at the Il Ciocco Conference Center near Lucca, Italy. It was chaired by Dr. Pierre Divenyi of the EBIRE Speech & Hearing Research Center in the U.S.A.

There were 103 attendees from 32 countries (Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, India, Iran, Ireland, Israel, Italy, Japan, Jordan, Kazakhstan, Netherlands, Portugal, Romania, Russia, Slovakia, Slovenia, Spain, Sweden, Turkey, Ukraine, United Kingdom, and the U.S.A.). There were 36 formal presentations (each about an hour and a quarter) over ten work days (and two free days) with a respectable amount of time for questions-and-answers and general discussion. It was the second such meeting - with the first (chaired by Dr. Steven Greenberg) held at the same location about four years earlier.

Each attendee received a 3-cm thick conference proceedings containing relevant recent papers by each of the speakers. There is a conference web site (at <http://www.ebire.org/earlab/asi2002.html> ). containing links to the web sites of each of the speakers. And there will be a book published containing submissions by the main speakers.

It is not my intent to review all presentations - but to concentrate on some worthwhile points made by a few. These points, in my opinion, best illustrate the issues being addressed and the proposed future research directions. Obviously, my personal interest in the signal-processing aspects of speech and hearing has colored my selection of presentations to highlight.

## Technical Focus<sup>1</sup>

Translating a message to be communicated into speech produces a series of changes in the speaker's vocal apparatus. These articulatory gestures, one after another, change the state of his vocal folds and vocal tract. These gestural changes in turn produce a complex acoustic signal which slowly varies in amplitude and spectrum. The listener must then "decode" this waveform into perceptual patterns from which he is able to infer the articulatory gestures that generated them. Speech communication therefore is based on the dynamics of both input and output - production and perception.

However, dynamics has not been the principle focus of the study of speech - perhaps because speech science has sprouted out of the tradition of phonetics and phonology - two fields historically preoccupied with isolated speech sounds (i.e., viewing speech as simply a string of phonemes). Actually such a string represents a sequence of intended targets of articulatory gestures, although we know that these targets need not be (and often are not) reached in order for the listener to recover the message. The relatively new area of spoken language processing has provided dramatic demonstrations of how much information a listener is able to recover from speech presented in remarkably high levels of noise and reverberation. Possibly one way he is able to do so is by relying on the patterns of change in the speech signal, which could act as acoustic pointers that are noticed by virtue of their motion.

The study of dynamic processes in speech has lead to a re-examination of fundamental questions in phonetics, linguistics, neuroscience, and speech technology. This NATO advanced study institute was designed to address these issues. Answering these questions should vastly increase our present understanding of the immensely complex process of speech communications. The answers also would be likely to provide clues for the development of more-natural speech synthesis schemes as well as for a much-hoped-for breakthrough in speaker-independent automatic recognition of noisy and reverberant speech.

## Defense Relevance

In its simplest form, a decision support system is basically a hierarchy of communications links between low-level "information gatherers" and the top-level "decision maker". In a convoy of Navy ships, for example, this hierarchy would stretch from various low-level subsystem operators

---

<sup>1</sup> This section is taken largely from the Preface of the conference proceedings (written by Dr. Pierre Divenyi).

(e.g., sonar operators, radar operators, aircraft pilots, weapons fire control stations) through various mid-level officers (e.g., intelligence analysts, battle planners), up to the commanding officer on the flagship. While these communications links frequently involve graphics and video, the overwhelmingly predominant mode of communications is speech - operators at consoles talking either across the room to the local junior officer in charge or into a microphone to other officers at remote locations throughout the fleet. Important engagement decisions are made on the basis of verbal communications progressing through this hierarchy. This communications structure works well most of the time. But events can occur which severely degrade this vital communications linkage at precisely the time that it is needed the most. For example,

- a.) In a real battle, this verbal communication takes place in a very noisy harsh environment. Sounds of explosions, low-flying aircraft, and emergency crews yelling orders can dominate the acoustic environment on which this decision-support information flow depends.
- b.) In a real battle, voice patterns deviate markedly from normal as high stress levels (or possibly panic) introduce an element that was not present during routine training.
- c.) In a real battle, casualties can occur and the manning of various nodes of this communications hierarchy will change. When an officer hears a strange unknown voice instead of the one expected, he might at first tune-it-out as a distraction before realizing that this new voice carries vital information.
- d.) A real battle is a very dynamic rapidly evolving environment, and unexpected events frequently occur. Verbal communications patterns then must necessarily deviate from those learned in training to accommodate new modes of response.

Yet despite the ubiquity of speech communications in decision support, very little research has been done on ways to improve the efficiency and reliability of such an intricate structure. Research needs to be performed on (a) how best to derive knowledge and interpretation of the tactical and operational pictures (e.g., Who is talking to whom? When? Where? How much? About what? etc.), (b) how best to sort, filter, and visualize all of this verbal communications as it relates to the common operational picture, and (c) how best to improve battlespace deconfliction through automatic platform identification, speaker identification, language identification, and topic spotting to rapidly determine threat versus friendly forces.

### **What Are the Essential Clues for Understanding Spoken Language?**

Dr. Steven Greenberg of the University of California at Berkeley in the U.S.A. is exploring the interface between sound and meaning (or at least lexical form). He started by making the observation that "speech is remarkably stable - even in a strong reverberation environment". He therefore hypothesized that the brain must be doing something that is impervious to different time delays at different frequencies. Thus, if we want to understand how the brain actually performs speech understanding, then analyzing this effect might give us further insight. His experiments involved taking speech recordings, performing quarter-octave filtering on them, discarding most of the channels produced, introducing relative time delays into different channels, reconstituting the sound, and asking subjects to assess their intelligibility. His results showed that syllable-length segmentation of speech seems to be more robust to this desynchronization than traditional phoneme-length segmentation.

He also pointed out that the peak of the modulation spectrum (to be described shortly) of speech decreases somewhat in a reverberant environment, but remains prominent. This suggests that the modulation spectrum might be a good representation of the intelligence or meaning in a sound stream and therefore be a good algorithmic tool for understanding this effect.

He went on to experiment with providing both visual and auditory inputs to the listener, and introducing relative time delays in each. He concluded that much better intelligibility was achieved when the video preceded the audio. Something in our audio-visual processing likes to see a visual input first with audio coming secondarily.

## **Time Frames of Spoken Language**

In Steve's second presentation, he explored the appropriate time lengths for analyzing spoken language. He started by pointing out that, in writing, there is a strong correlation between word length and word frequency (i.e., short words are used more frequently) - whereas in speech, this is not the case. A plot of frequency-of-occurrence vs. word length for a large body of typical speech has just one peak with tails on either side. He went on to show that, on the other hand, "stress accent" is closely related to word length. A plot of frequency-of-occurrence vs. word length typically shows two peaks - a short-word-length peak for unaccented words and a long-word-length peak for heavily accented words.

In the discussion period following Steve's presentation, the ensuing debate as to whether frequency or time is more important reminded me of the "Tastes great vs. less filling" light beer commercial.

## **Neural Dynamics of Speech Perception**

Dr. Stephen Grossberg of Boston University in the U.S.A. is internationally renowned for his pioneering work in the neurological models of perception - leading to his classic papers on adaptive resonance theory in the 1970s. He described the following experiment: A subject listens to repeated recordings of the form

"(broadband noise) eel was on the (word)"

If the word inserted in the recording was "axle" then the subject reported hearing "wheel was on the axle". If the word inserted was "shoe" then the subject reported hearing "heel was on the shoe". If the word inserted was "orange" then the subject reported hearing "peel was on the orange". If the word inserted was "table" then the subject reported hearing "meal was on the table". That is, the brain's perception of the broadband noise depends upon the word used at the end of the phrase. This is an example of the future effecting the past - or of meaning effecting phonetics. In other words, when the brain is presented with the context of a meaningful sentence, it interprets the noise according to the meaning of the context.

Another observation was that human speech intelligibility is relatively immune to fast vs. slow speeds. The brain seems to adapt the silence segments to have a mean length. On the other hand, there seems to be something important about the length 150 milliseconds. This figure on inter-segment duration keeps popping up over and over again as an inflection point in the results of different experiments and in different contexts.

He gave several other experiments involving sounds and perception. Most of these could be explained by his 1997 hypothesis that the human auditory filter was actually two parallel filters - one processing consonant transients, and the other processing steady-state sustained vowels. That is, we have two different working memories, with one acting as an automatic gain control on the other.

## **Evidence for Multi-Resolution Analysis of Auditory Stimuli**

Dr. George Meyer of Keele University in the United Kingdom observed that models of speech perception make explicit or implicit assumptions about the order of processing and the representations that are being processed. Some models assume very fine-grained analysis of the speech signal to extract features such as the voice pitch and speaker localization, while others are based on much coarser representations, such as formant structure. There are considerable differences between these models, and these differences are not easy to reconcile because perceptual data appears to support both views. He argues that different tasks require different representations and that speech processing is carried out simultaneously at different resolutions.

He described an experiment in which a chirp wave segment was inserted into the second formant of a synthesized recording of a vowel-to-nasal syllable transition. The presence of the chirp strongly affected which nasal sound was heard. The fact that the subjects could easily hear the chirp is evidence for a high-resolution analysis. The fact that the chirp changes the perceived category of the nasal suggests that the pattern matching is carried out in a low-resolution representation.

## **Acoustic/Modulation Frequency Transforms for Single-Channel Talker Separation**

Prof. Les Atlas of the University of Washington in the U.S.A. explored the case of listening to simultaneous talkers. He pointed out that humans are highly insensitive to talker overlap, yet most automatic speech recognition systems have great difficulty with this. What is it that humans are doing that makes it so easy for us to listen to two or more simultaneous speakers? He described recent auditory physiological evidence pointing to a modulation frequency dimension in the auditory cortex - existing jointly with the usual acoustic frequency dimension. This leads to the so-called "modulation frequency transform" - a relatively slowly-varying two-dimensional representation of sound wherein the first dimension is well-known acoustic frequency and the second dimension shows the modulation imposed on each frequency channel of the first dimension.

Les is developing a formalism for this transform and he urged all conference participants to help him in attempting to standardize on terminology, parameter specifications, and measures-of-effectiveness. He went on to demonstrate results of applying this transform to recordings of simultaneous talkers - producing a visual two-dimensional color map. The audience could clearly see a separation of multiple talkers in pitch and formants. Since the transform is invertible, identifying and masking out the obvious information from the undesired speaker, followed by the inverse transform results in the removal of sonorant information such as vowels and voiced consonants from the undesired talker while preserving the speech of the desired talker.

He explained the various ways to introduce the requisite non-linear operation between the first and second transform. He also explained why he thought that the traditional magnitude-and-phase variables associated with the Fourier transform lead to much confusion in their application to speech processing, and recommended the adoption of the scale-and-shift variables associated with wavelet transforms.

He went on to show preliminary results indicating that an audio codec could be designed using these modulation transform principles yielding significantly better performance (lower bit rates) than current approaches.

I speculate that we will be seeing a lot more of the modulation frequency transform in the years ahead as its usefulness to other acoustic (and even non-acoustic) applications is explored.

## **The Auditory Image Model and its Use in Speech Analysis Time-Domain Auditory Processing of the Dynamic Aspects of Speech**

To me, the two presentations by Prof. Roy Patterson of Cambridge University in the United Kingdom were the highlight of the conference.

He speculated that millions of years ago, early life forms (e.g., shrimp) learned to communicate by banging together whatever bony appendages they had to produce "clicking" sounds. The resonance properties of these clicks contained useful information such as the approximate size of the originator. Soon these life forms learned that communications would be better if a sequence of multiple clicks were generated (since the listener might miss a single click). Millions of years later, modern man uses the frequency of these clicks to produce the pitch or fundamental tone of his talking or singing, and modulates the resonance structure (formants) associated with each pulse through his vocal tract. Again, this leads naturally to the use of the modulation-frequency transform as a useful tool for analyzing sound (particularly speech). In humans, we know that the frequency transform part of this operation is done by the cochlea. Roy reasons that the modulation frequency part (for which he uses a time-interval transform) is performed in the brain stem. He speculates that a "scale transform" is produced in the higher levels of our auditory system producing what he calls an "auditory image" for analysis and categorization.

He showed several graphical examples (similar in concept to those of Les Atlas above) of the information gained by using the modulation-frequency transform on speech and musical data.

## **Assessment**

This conference was extremely useful in bringing together experts from diverse backgrounds and facilitating their interaction to explore this new field of study. In addition to the expected speech and hearing experts present, scientists representing the fields of human physiology, cognitive science, neuroscience, psychology, and computing science were actively involved. While much time was spent in arguing over semantics as experts within each field tried to reconcile their terminology with that of others, I believe that this was necessary and beneficial as they struggled with specifying the issues and formulating their research directions.

I greatly enjoyed the conference and was very impressed with the high technical calibre of the leaders, presenters, and attendees. As usual, a great deal of very useful information exchange took place over meals and coffee.

## **Contacts**

Dr. Pierre Divenyi  
EBIRE Speech & Hearing Research  
VA Medical Center  
150 Muir Road  
Martinez, CA 94553  
USA  
email: [pdivenyi@ebire.org](mailto:pdivenyi@ebire.org)

Dr. Keith Bromley  
Associate Director for Information Technology  
Office of Naval Research International Field Office  
London, United Kingdom  
email: [kbromley@onrifo.navy.mil](mailto:kbromley@onrifo.navy.mil)

The Office of Naval Research International Field Office is dedicated to providing current information on global science and technology developments. Our World Wide Web home page contains information about international activities, conferences, and newsletters. The opinions and assessments in this report are solely those of the authors and do not necessarily reflect official U.S. Government, U.S. Navy or ONRIFO positions.

[Return to ONRIFO Newsletters](#)