BEYOND CLOSE-TALK – ISSUES IN DISTANT SPEECH ACQUISTION, CONDITIONING CLASSIFICATION, AND RECOGNITION

Vincent Stanford, Cedrick Rochet, Martial Michel, John Garofolo
Stanford@nist.gov, Cedrick.Rochet@nist.gov, John.Garofolo@nist.gov
US National Institute of Standards and Technology
Martial.Michel@nist.gov
Systems Plus, Rockville MD

ABSTRACT

Properly designed reference data and performance metrics can offer crucial aid to developers of advanced statistical recognition technologies. We focus here on audio data acquisition from close-talk, nearfield, and farfield sensors, and upon its processing, and its metrology. Our intention is to support the research community as it develops state of the art data acquisition and multimodal processing algorithms by supplying standard reference data, metrics, and sharable infrastructure, rather than developing the algorithms themselves.

1. OVERVIEW

Most contemporary commercial speech recognizers were designed to operate with close-talk microphones that provide very high signal-to-noise ratios; and they perform best in quiet environments. However, meeting spaces contain numerous people, who move as they present ideas; may resist being tethered to microphones; and often speak simultaneously. Therefore, multi-sensor speech acquisition and phased array processing may offer productive avenues for research in multimodal meeting interfaces.

Historically, precision metrology of speech recognizer performance allowed system developers to assess the impact of proposed algorithm changes on actual system performance, and hence was critical to the development of more effective technologies. For example, precision Word Error Rate measurements [1] allowed speech recognizer developers to compare performance of numerous processing algorithms such as LPC vs. FFT Mel cepstra, Viterbi versus stack decoders, and bigram versus trigram language models. Armed with this precision performance measurement tool, and with standard reference data, they could proceed confidently with incremental upgrades.

In order to support new research initiatives in multimodal systems, the NIST Meeting Room and Smart Space Projects are making reference data, data acquisition infrastructure, and additional measurement tools available to facilitate research and development in this area. For example, the technical objective of our third generation microphone array, the NIST Mk-III, is to support speech corpus development, and also to provide interested laboratories access to a relatively cheap and reliable means of acquiring multi channel speech signals suitable for phased array processing research.

Further, our meeting room audiovisual data corpus offers multiple views of the same speech, at multiple distances ranging from a few centimeters at head mounted microphones, to perhaps a few tens of centimeters at lapel microphones, to one or two meters at tabletop microphones, to several meters at medium field or farfield microphone arrays. We believe that this multi-distance capture is presently unique in a publicly available speech corpus, and that it can facilitate research on algorithms to optimally estimate close talk speech from more distant sources, quantify the changes due to distance, or study of source separation. The simultaneously captured video views may also allow for sensor fusion experiments.

2. ISSUES IN MULTISENSOR MEETING SPEECH ACQUISITON

In a communication to the authors, one array signal-processing expert commented wryly that microphone arrays would not cause cepstral coefficients to form on the lips of the speakers in a room and jump to the decoder stage of a speech recognizer. [2] This comment stimulated a discussion on the preprocessing and algorithmic areas that would be needed for the effective use of distant microphones, either singly, multiply, or in geometrically coherent arrays in meeting rooms. It was clear that there are several new challenges that will be encountered by investigators attempting to process speech from meetings. In order to draw the community into this discussion, we present a preliminary list these below. While we are certain this list is not complete, these issues will clearly present technical challenges and opportunities to the research community.

Multiple speakers – The very nature of meetings means that there will be multiple speakers in each meeting. The present state of the art of large vocabulary speech recognition supports high accuracy only by speaker adaptation, or speaker dependent training. Therefore there will be an opportunity for improved systems that segment the speech according to the individual speakers, and use speaker de-

pendent acoustic models, or adapt to the particular speaker. Segmentation by speaker will also be essential to the production of rich meeting transcripts.

Overlapping speech - Speech recognition system performance in meeting room environments is challenged by the prevalence of overlapping speech. Approximately seventy percent of the speech segments of the NIST Meeting Room Pilot Corpus contained overlapping speech. Moreover, even the close-talk microphones show significant bleed through from multiple simultaneous speakers; and the tabletop and array microphones contain all of the speakers, which must be dealt with by technical means.

Multi-distance views of speech – There are many sensors in an instrumented meeting room including close talk, lapel, tabletop, and more distant array microphones. Overlapping speech removed by algorithmic means if the speech from each speaker is to be decoded with high accuracy. A somewhat oversimplified acoustic computation yields a variation in sound intensity inversely proportional to the squatted distance from the source. This implies approximately a 60 dB variation in amplitude due to the drop from close talk to microphone array distances. This suggests that the traditional 16 bit ADCs used for speech recognition will provide insufficient dynamic range; so modern ADCs with twenty-four bits should be used for multi-distance views of speech.

Synchronization of multi-channel databases – In a meeting room instrumented with multiple sensors, running on commodity hardware, capture of numerous sensor streams naturally occurs asynchronously. There are variations in the system and capture device clocks; and Windows, Linux, and Unix platforms are not real time systems, so the time tags that can be obtained are perturbed by both drift and random delays. Statistical corrections are needed to bring the data streams into a usable degree of synchronization for audiovisual processing. More stringent requirements obtain for audio phased array processing.

Speech detection - Speech fricatives and plosives will be hard to detect in the farfield, so reliable speech onset times may be difficult to obtain.

3. TECHNIQUES FOR DISTANT MULTISENSOR SPEECH ACQUISITON

There is a vast literature on algorithms for processing signals acquired from distant microphones in the near and far fields and many excellent review sources are available such as [3], [4], and [5]. Our purpose here is to simply note how some of these techniques could be studied using the NIST Meeting Room corpus, and the associated metrology requirements. Some of the major areas addressed are given below:

Beamforming - Various beamforming techniques have been developed and extensively studied, but because speech is relatively broadband, with fricatives having energy beyond 10 kHz and pitch frequencies down to the 100 Hz range, narrowband techniques are not recommended. The NIST Meeting room offers multiple sources of speech data. The variety of sources offers research opportunities in nearfield techniques, farfield, and interarray beamforming.

Source location – Various techniques for source localization include subspace methods like MUSIC, and ESPRIT, as well as broadband techniques like TDOA. The possibility to investigate audiovisual sensor fusions will also be supported by the Meeting Room corpus; for example allowing face localization in video to be combined with acoustic DOA estimates.

Autodirective speech acquisition – If speakers can be localized, from distant microphone arrays, either of random or regular geometries, then the close talk speech in the Meeting Room corpus can be used to test and refine algorithms to estimate the speech near the mouth.

Echo cancellation – Adaptive filtering techniques can be applied to multiple channels that have cross contamination and estimate the transfer functions between the channels.

Channel normalization – Cepstral mean normalization, was required for microphone independence in first generation speech recognition systems. The channel mismatches from speech obtained at close talk, nearfield, and farfield are even more severe, and may require new techniques. Phased array processing may also cause a channel mismatch between beams formed at various bearing angles. Also, high frequencies are differentially lost with distance.

Inter array processing – Multiple NIST Mk-III microphone arrays can be slaved to a master and use a common clock signal from the master board. We have tested this with up to four arrays with 256 microphones and are interested in community input on whether we should actually synchronize these arrays to a single clock for the Meeting Room data collection.

Speaker identification – The present state of the art in large vocabulary speech recognition allows very high accuracy for only for well-known speakers after user specific training, or adaptation. The Meeting Room data will allow experiments in speaker identification and segmentation, which could ultimately allow for speaker dependent or adaptive methods to be employed for meeting transcription.

Pre- and post-filtering – Without pre-filtering, we found that adaptive beamforming weights would not converge in

our reverberant lab spaces, apparently due to too much low frequency energy that could not be steered out because its wavelength exceeded the diameter of the array. Applying a bandpass filter, with a half-power-points at 190 Hz and 7,200 Hz, to each array channel before adaptive beamforming allowed the weight matrix to converge. More advanced pre- and post-filtering could allow improved signal estimation and subsequent recognition rates.

4. INFRASTRUCTURE AND TOOLS

We have developed a number of tools including speech SNR metric, data transport, distributed processing, and annotated reference data in support of our Meeting Room project, which are being adopted by several independent laboratories. We used the NIST Data Flow System to implement a spoken language data acquisition facility for the NIST Meeting Room project. It will have 280 microphones in four Mk-III microphone arrays, twenty-four random placement microphones, seven HDTV resolution cameras, and a smart whiteboard. These generate almost two gigabytes of data per minute, which we time tag to video frame resolution for the speech and video research communities.

4.1 NIST SPEECH SNR MEASUREMENT

The traditional speech signal to noise measurement technique imputes a single SNR level to the speech. An analysis of the amplitude distribution of speech mixed with background noise shows that the distribution is not even approximately Gaussian in form. We found that for speech in noise the amplitude is quite well represented by a tri-Gaussian mixture with common mean. The three standard deviations represent a best fit to the background noise, unvoiced speech, and voiced speech, which yields two distinct SNR levels, σ_u, and σ_v rather than the traditional single measurement. We make this measurement algorithm available in our open source project.

4.2 NIST DATA FLOW SYSTEM: A CONNECTIVITY TOOLKIT

The NIST Smart Data Flow System was developed to provide connectivity to the large number of sensors and devices needed to acquire the reference data for meeting recognition research. Figure 1 shows an operational flow graph for data review in the NIST Meeting Room. The NIST Data Flow System generates the data flows and transports the data among the distributed clients. The system consists of a defined middleware API for real-time data transport, and a connection server for data sources, processing, display, and storage clients. Hand crafting the needed inter-process communication was found to be very labor intensive and brittle with respect to changing requirements for new sensors and configuration changes forced by equipment faults.

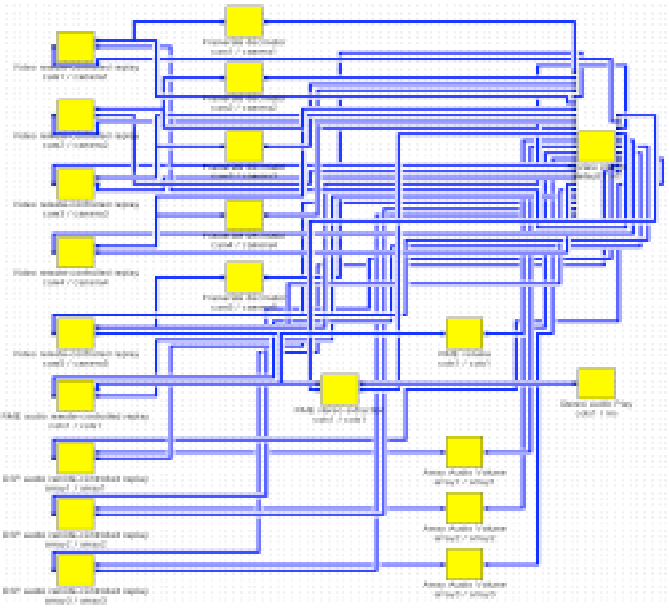

Fig-1. Executable Smart Data Flow System graph for Meeting Room data processing and review.

The Smart Data Flow System toolkit has components for graphical configuration of flow graphs, allocation of the graph nodes to distributed systems, and connection via TCP/IP. The data transport code is provided in the NIST Data Flow System libraries. It also has a code generator for a simplified construction of clients that operate as nodes of data flow system graphs. This can very substantially reduce the systems programming burden in research and development laboratories working on multi modal sensor based interfaces, or advanced real time classification systems.

4.3 THE NIST MARK-III MICROPHONE ARRAY

In the context of our meeting speech acquisition program we have developed the third generation microphone array, which has been of general interest to multimodal research programs. Design goals included placing the preamplifier and analog-to-digital converter very close to the microphones to reduce analog noise from the environment. We also wanted to be independent of particular DSP cards, which can become unavailable unexpectedly, or computer hardware or operating systems. This required placing the microphones on daughter boards, which we called the Microboard, with the preamplifiers, and analog to digital converters. The second board, the Motherboard, reads the serial outputs of the eight daughter cards under control of an FPGA with a VHDL program, as shown in Figure-2.

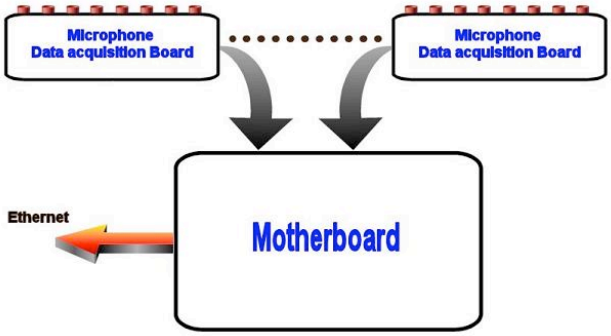

Fig-2. Mark-III Microphone Array overview showing microphone daughter cards and the networked motherboard

4.3.1 MARK-III ARRAY DESIGN DETAILS

The Microboard implements three basic processing stages of data acquisition including:

- Amplification, using a double stage operational amplifier
- Analog to digital conversion with a stereo 24-bit multi rate converter up to 96 kilo-samples
- Serial connection to the motherboard

The microphones are spaced in 2cm. intervals, so the diameter of a 64-microphone array is 128cm. This value is chosen to process voice frequencies up to 8Khz. without spatial aliasing. Another version of this board with a 4cm. separation provides a lower frequency range but higher directional resolution at low frequencies. It also allows octave-nested arrays to be constructed, as first described by Flanagan [6], to extend the diameter of acoustic arrays while maintaining satisfactory directional resolution with a fixed number of sensors. Figure-3 shows a Microboard populated with microphones and integrated circuits. The Electret microphones shown are available for about fifty cents US each and contribute to the low cost of the design.

The figure-4 shows the Motherboard, which is the data-gathering platform for the eight Microboards through the connectors shown at the bottom. The logic for data gathering is executed in a Field Programmable Gate Array (FPGA), which provides more useable I/O pins and bandwidth than competing microcontrollers available during the design phase. It also has support logic to provide:

- Four Mbytes of SRAM data buffer
- Fast Ethernet Physical Layer Device (PHY)
- DIPswitch to configure the MAC address
- A clock synchronization signal to propagate the master board clocks to other slave microphone arrays
- PROM containing the firmware
- Condition indicator LEDs

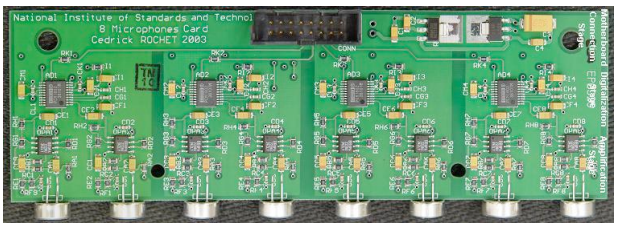

Fig-3. A Microboard is shown here with eight microphones, eight preamplifiers, four stereo 24-bit ADC chips and serial connector.

The VHDL logic on the board interfaces the different parts of the Motherboard in order to make them fully operational as a stand-alone networked sensor system. These include:

- Ethernet interface: a stack of Ethernet protocols was implemented in the chip in order to communicate across the network with any computer through the Physical Layer Device and then the Ethernet,
- Memory interface: on board memory enables about half a second of live data to be buffered (2Mbytes for 64 channels at 22050Hz) if UDP packets are missed,
- Microboard interface: a double buffering system to gather and reorder the serial data coming from the 8 Microboards each with eight microphones.

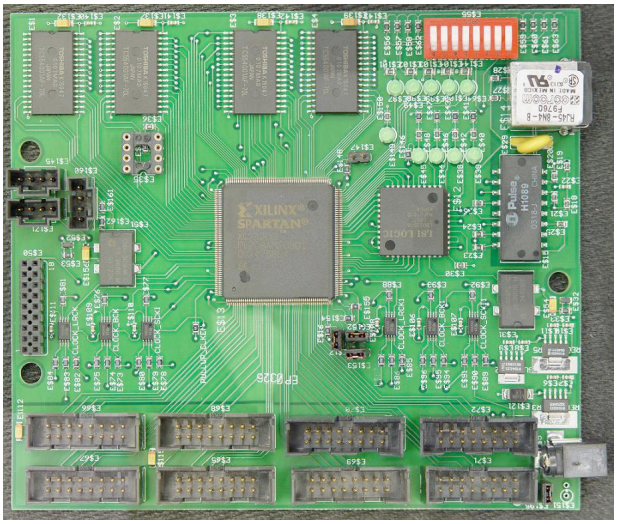

Fig-4. A Motherboard is shown here with the FPGA, SRAM data buffer, and serial connections for Microboards.

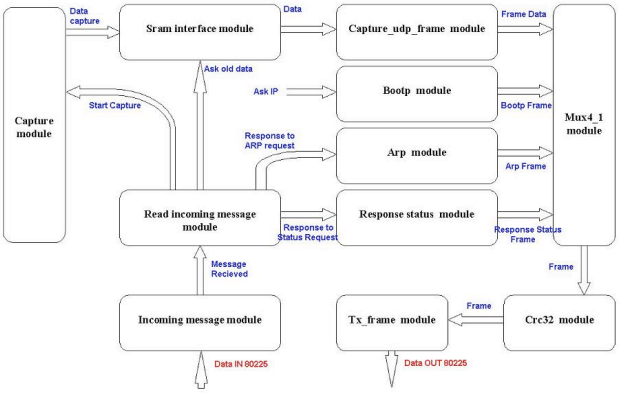

Fig-5. Block diagram of VHDL data acquisition and distribution software resident in the FPGA

4.3.2. A NETWORKED SENSOR ARCHITECTURE

With power and Ethernet cable, the Mk-III Microphone Array broadcasts UDP packets to find a BOOTP server on the network to get an IP address. Then any computer on the network can interact with it through software developed at NIST and freely available at our website (**www.nist.gov/smartspace**/toolchest). Our smart space open source project offers a range of software utilities from a 64 channel digital oscilloscope to java clients to read the array data and save it to disk.

The amount of data is about 4.4MB/s when the capture is done at a frequency of 22,050Hz and about 9MB/s at a frequency of 44,100Hz. Due to the high number of UDP packets sent, and the fact that UDP is not a reliable protocol, a PC on a local switch should be dedicated to data collection and transmission to the surrounding NDFS clients for processing. To avoid packet loss, on the Linux operating system we use for data acquisition, it may be necessary to tune the kernel through sysctl.conf in order to increase TCP buffer limits, or upgrade the hardware to guarantee enough speed to avoid missing packets.

5. CONCLUSIONS

The objective of the National Institute of Standards and Technology is to aid industry in competitiveness and productivity through measurements and standards. In aid of this goal we create and provide spoken language reference data for training and test of recognition algorithms, as well as tools for data acquisition, quality measurement, and performance metrology.

We have also reviewed the NIST Smart Space project data flow infrastructure and test bed for integration of advanced systems, such as the Microphone Array Mark III, and other audiovisual sensor systems. Also, the NIST Meeting Room project gathers archival data from multiple arrays as reference data for numerous research efforts.

We are currently exploring the possibilities of standards working groups on interoperability, and interested parties are encouraged to contact the authors by e-mail at NIST.

6. DISCLAIMER & LICENSE STATEMENT REGARDING THE SMART DATA FLOW AND MICROPHONE ARRAY MARK III

Employees of the Federal Government in the course of their official duties developed the described software and hardware at NIST. Pursuant to title 17 Section 105 of the United States Code this software is not subject to copyright protection and is in the public domain.

Certain commercial products have been identified in order to adequately specify or describe the subject matter of this work. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for any particular purpose.

The NIST Data Flow System and The Mark-III Microphone Array are experimental research systems. NIST assumes no responsibility whatsoever for its use by other parties, and makes no guarantees, expressed or implied, about their quality, reliability, or any other characteristic.

NIST and the Smart Space project would appreciate acknowledgements if these tools are used for research purposes.

REFERENCES

[1] W. Fisher, and J. Fiscus, *Better Alignment Procedures for Speech Recognition Evaluation,* ICASSP 1993, Vol. II.

[2] J. McDonough, *Unpublished communication to the first author*, circa September 2003.

[3] D. Johnson and D. Dugeon, *Array Signal Processing*, Prentice-Hall, Inc. Englewood Cliffs New Jersey 1993.

[4] M. Brandstein and D. Ward, *Microphone Arrays Signal Processing Techniques and Applications*, Springer-Verlag, Berlin Heidelberg New York, 2001.

[5] H. Van Trees, Optimum Array Processing, John Wiley & Sons

[6] J. Flanagan, D. Berkley, G. Elko, J. West, M. Sondhi, *Autodirective Microphone Systems*, J. Flanagan, D. Berkley, G. Elko, J. West, Acustica, Vol. 73, 1991 p. 58-71