Getting Two Birds with One Phone:
An acoustic sensor for both speech recognition and medical monitoring

Presented in poster format at the 138th Meeting of the Acoustical Society of America, Columbus, Ohio
2 November 1999

James D. Bass, Michael V. Scanlon, Thomas K. Mills, ARL
John J. Morgan, USMA

Army Research Laboratory
2800 Powder Mill Road
Adelphi, MD 20783
jbass@arl.army.mil or mscanlon@arl.army.mil

ABSTRACT

Automatic speech recognition (ASR) in noisy environments requires innovative use of disparate technology to overcome the special demands caused by multiple speakers and minimal Signal to Noise Ratios (SNR). Use of non-airborne acoustic sensors (physiological sensors) for ASR imposes special requirements on speech engines due to the changes in spectral information caused by alternative pickup locations. Specifically, the relative powers of voiced and non-voiced components are often reversed when compared to relative powers of these components collected with conventional microphone technology. Our research entails the evaluation of various ASR sampling configurations in conjunction with different body location points for the physiological sensor. Sensor detectable physiology provides clues to a user's stress, cognition, and general health, thus our physiological sensor package supports the ability to perform speech recognition and medical monitoring (two capabilities for the price of one sensor). For ASR, our goal is to build a suite of optimal sampling configurations for several strategic body locations (e.g., throat, temple, thorax, etc.). The experimental design includes traditional word error rate and subjective task completion components. These experiments were conducted in environments with speech SNR ranges of 10, 3, and 0, decibels. These speech SNR ranges cover the acceptable commercial ASR environment of 10dB to ranges where commercial ASR systems with conventional microphone technology are completely ineffective. Our pilot study indicated excellent performance for a sensor that is throat located.

1. Introduction

ARL has developed sensor technology to monitor a soldier's voice and physiology by using enhanced acoustic sensors. The sensors consist of a fluid or gel contained within a small, conformable, rubber bladder or pad that also includes a hydrophone. This enables the collection of high signal-to-noise ratio (SNR) cardiac, respiratory, voice, and other physiological data. The pad also minimizes interference from ambient noise because it couples poorly with airborne noise. It is low cost and comfortable to wear for extended periods.

ARL further developed (with support from the United States Military Academy), a Hidden-Markov Model (HMM) phoneme-level speech system optimized for the physiological sensor package. Lab observations indicated tremendous potential for use in automatic speech recognition under quiet and very high noise conditions. The next step was to develop an experimental design to test the effectiveness of the sensor in a more controlled setting.

2. Experimental design (the pilot study)

2.1. Training

Our investigation considered two kinds of recognizers. Those trained on data collected via the physiological contact sensor (physiological) and those trained on data collected via the conventional airborne microphone (airborne) and two kinds of speech waveform data, data input via the physiological and data input through the airborne microphone. It is reasonable to assume that under any conditions, airborne recognizers will work better on airborne data than physiological recognizers, and physiological recognizers will work better on physiological data than airborne recognizers. We claim however, that in environments with high levels of ambient noise, error rates for physiological recognizers will always be lower than for airborne recognizers. To show this, several separate recognizers should be developed. We have so far trained two recognizers, one exclusively on physiological data and the other exclusively on airborne data. Both of these sets of training data were collected in a quiet studio-like environment. We considered the following variables for recognizer training: input device and speech SNR levels. The input data streams take on three values: physiological, airborne, and both airborne and physiological. The noise level variable would ideally take on a continuum, but for our current project we focus on three values 0-, 3-, and 10dBA SNR. Follow-on experiments could introduce more sophisticated acoustic models, such as phonetic granularity that would take on the values of mono, bi, or triphones.

2.2. The Corpus

Training a speech recognizer requires a large database of speech signals. Such a database is referred to as a corpus. Our corpus was collected by, and composed of speech samples from, members of the Department of Foreign Languages at the United States Military Academy (USMA) with the following conditions. The donors to the corpus also known as informants, were asked to read from a prepared set of 50 phrases. To sharpen our initial results we restricted our set of informants to male native English speakers. The 35 informants donated their speech to our corpus. Each informant read the script 5 times under 4 different ambient noise levels while wearing the two sensors. The physiological sensor was strapped on to the right side of the throat, and a head-mounted airborne microphone was positioned to the front of the mouth via a boom configuration. Each training subject read the script twice in quiet conditions. Then the speech SNR of a typical speaker in the recording booth was set sequentially at 10-, 3-, and 0-dB by repeatedly playing a spectrally stable section of a recording from inside the turret of an M1A1 tank traveling at 25 kph. Both sensors were sampled at 44.1 kHz and quantized to 16 bits by a standard PC soundcard.

2.3. The Models

All recognizers developed in this experiment were based on Hidden Markov Models (HMM). We used Entropic's HTK for development, training, and testing of the models. We extracted from the waveform data 39 dimensional parameter vectors consisting of 12 Mel cepstral coefficients, 1 energy component, 13 first order regression components calculated from the first 13 components, and similarly 13 second order regression components calculated from the second 13 components. Each emitting state of the HMMs has a corresponding 39 dimensional mean vector and a 39 by 39 diagonal covariance matrix obtained from the parameter vectors. These statistics characterize a gaussian distribution that models the generation of the observed speech data. The transitions between states in the HMMs are probabilistic, so each HMM has an upper triangular transition matrix whose entries are calculated from state transition occurrence counts.

2.4. The Phonetic Transcription

Before transcribing speech into phonetic segments, a decision must be made concerning the granularity of the segments that will be used in the transcription. The granularity can range from a broad transcription that only uses two symbols v and c to denote vowels and consonants to a narrow phonetic transcription that uses close to 100 symbols and includes transcriptions of breathing sounds in addition to phonemes and all phones. In other words, the transcriber must decide on a list of symbols that he will use in transcribing the speech. There are several lists of phonetic units that have been widely used in speech recognition research, e.g. the DARPABET, TIMIT, CMU, etc. We chose to use the CMU phone list in this pilot study, because we have access to the CMU pronouncing dictionary and thus we saved ourselves considerable effort by not having to phonetically transcribe most of the words in our vocabulary. However, we performed some modifications to the CMU phone list and its dictionary for our purposes. We replaced dd, kd, pd, and td, by d, k, p, and t respectively. We added three words that were in our data collection script but were missing from the CMU dictionary. For each phone in the resulting list of 41 phones we made one monophone which is a standard 5-state (only 3 emitting), HMM.

2.5. The Training Strategy

The first 21 subjects (designated trainers) were used to build the speaker independent HMM models for both the physiological and airborne sensors. The remaining 14 subjects (designated testers) were reserved for actual tests of the models and performed no training. We used a flat start and Baum Welch embedded re-estimation training strategy. To bootstrap the Baum Welch embedded re-estimation training algorithm we initialized the HMMs with the global covariance matrix. The global covariance matrix is calculated over all of the data and stored in each of the HMMs. After a couple of training iterations a short pause t-model was inserted between words and training continued. Then we used forced alignment to allow the recognizer to choose between the silence and short pause models between words. Finally, we gradually increased the number of gaussians up to 28 mixtures.

2.6. The Language Model

Since we want to measure the performance of the sensors as front-ends of the ASR engines, and not that of the HMM models themselves, we employed acoustic and language models that were simple in design but challenging for the models. In later experiments we intend to focus on optimizing the performance of the system as a whole by introducing more sophisticated modeling techniques like triphone acoustic models, statistical language models, and larger vocabularies. But for this pilot study, we used monophones for our acoustic models and simple sentence and word loops for our language models. Our first experiments were performed on sentence loops. By “sentence loop” we mean a language model that restricts the possible recognized sentences to the list of sentences in the corpus. In other words, the recognizer only “knows” the 50 sentences in the corpus and selects the most probable. After generating 1,000,000 sentences, HTK measured the empirical perplexity (average word branching factor) of this language model at 2.5, this is considered a simple grammar (a complex full grammar network is computed to be approximately nine). A word loop is a language model in which every word in the vocabulary can follow any other word. Here “vocabulary” refers to the list of the 153 unique words that appear in the corpus. The longest sentence in our set of prompts has ten words, so we constrain the language model to only generate sentences of at most ten words in length. The empirical perplexity of our language model after generating 1,000,000 sentences was measured by HTK at 183.9 (high, because all words branch).

3. Results

The tables below illustrate the relative performance of each sensor. Note that this data represents the results of the 14 test subjects using the speaker independent models.

Sentence Loop Model [% Perfect Sentence Recognition]		
dB	**AIRBORNE**	**PHYSIOLOGICAL**
0	40.6	96.5
3	60.7	98.7
10	98.7	99.4

Table 3.1. Sentence Loop Language Model

Word Loop Model [%Accuracy]		
dB	**AIRBORNE**	**PHYSIOLOGICAL**
0	-0.1	39.6
3	12.8	50.5
10	51.5	66.8

Table 3.2. Word Loop Language Model

The columns headed AIRBORNE contains error rates for the airborne recognizer and airborne test data, while the columns headed PHYSIOLOGICAL contains word error rates for the PHYSIOLOGICAL trained recognizer and physiological test data.

Table 3.3. Complete Test Subject Data

This table provides the test results for the 14 test subjects as evaluated by two commonly accepted word-accuracy-rate measurements.

Pilot Results for Airborne Sensor Using Airborne HMM						
	S/N = 10 dBA		**S/N = 3 dBA**		**S/N = 0 dBA**	
Informant	**%Corr**	**%Acc**	**%Corr**	**%Acc**	**%Corr**	**%Acc**
2200a	97.5	97.5	66.3	60.1	55.1	44.1
5959a	100.0	100.0	60.4	56.7	40.4	31.4
5601a	100.0	100.0	62.5	55.1	26.5	14.7
5452a	100.0	100.0	60.0	53.1	24.1	15.1
2955a	94.3	91.8	15.9	7.8	17.6	10.6
2852a	97.6	97.6	47.4	38.0	23.7	14.3
2655a	98.4	98.4	79.1	74.8	69.8	63.3
5816a	100.0	100.0	62.0	55.9	52.3	46.0
4625a	98.4	98.4	90.5	88.0	71.1	66.2
3205a	100.0	100.0	78.8	77.1	65.7	60.0
2854a	100.0	100.0	82.6	78.7	58.7	51.9
3113a	100.0	100.0	92.7	89.8	82.5	79.6
4030a	100.0	100.0	62.7	53.6	40.9	30.8
5203a	100.0	100.0	93.9	89.0	70.6	60.1
Mean	**99.0**	**98.8**	**68.2**	**62.7**	**49.9**	**42.0**
Standard Deviation	**1.7**	**2.2**	**20.7**	**22.6**	**21.1**	**22.7**

Pilot Results for Physiological Sensor Using Physiological HMM						
	S/N = 10 dBA		**S/N = 3 dBA**		**S/N = 0 dBA**	
Informant	**%Corr**	**%Acc**	**%Corr**	**%Acc**	**%Corr**	**%Acc**
2200c	100.0	100.0	100.0	100.0	100.0	100.0
5959c	100.0	100.0	100.0	100.0	95.9	95.9
5601c	100.0	100.0	100.0	100.0	100.0	100.0
5452c	100.0	100.0	100.0	100.0	100.0	100.0
2955c	100.0	100.0	100.0	100.0	95.9	95.5
2852c	95.5	95.5	98.4	98.4	98.0	97.1
2655c	99.6	99.6	100.0	100.0	99.6	99.6
5816c	100.0	100.0	100.0	100.0	99.6	99.6
4625c	100.0	100.0	98.4	98.4	94.7	92.7
3205c	100.0	100.0	100.0	100.0	100.0	100.0
2854c	100.0	100.0	97.1	97.1	87.4	86.5
3113c	100.0	100.0	100.0	100.0	100.0	100.0
4030c	97.1	97.1	91.0	89.8	90.2	87.4
5203c	100.0	100.0	100.0	100.0	100.0	100.0
Mean	**99.4**	**99.4**	**98.9**	**98.8**	**97.2**	**96.7**
Standard Deviation	**1.4**	**1.4**	**2.4**	**2.8**	**4.1**	**4.7**

$\%Correct = \frac{N-D-S}{N} \times 100\%$ $\%Accuracy = \frac{N-D-S-I}{N} \times 100\%$ Where N is sample size, D is deletions, S is substitutions, and I is insertions.

% Perfect Sentence Recognition: Measures the percentage of whole sentences completely and correctly recognized (no individual word errors).

4. Summary

The physiological sensor demonstrates clear and convincing superiority over the airborne acoustic sensor when exposed to noisy environments at and below 3dB SNR. We feel confident of this performance trend will be sustained in other types of noise. This pilot study strongly suggests the need for further

experiments using the physiological sensor with larger, more diverse subjects, and under more extreme and varied noise types. Scope will also expand to include multi-modal (gesture, gaze, lip reading, etc.) input devices and data fusion paradigms.

BIBLIOGRAPHY

Guan, C.T., Leung, S.H., and Lau, W.H.: "Multi-model approach for noisy speech recognition", *IEEE Electronic Letters*, 34(1), January 1998.

Chen, Tsuhan (editor): "The Past, Present, and Future of Speech Processing", *IEEE Signal Processing Magazine*, 15(3), May 1998.

Hansen, John H. L., "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communications*, Nov. 1996, Vol. 20, pp. 151-173.

Jelinek, Frederick: Statistical Methods for Speech Recognition, Cambridge, MA: The MIT Press, 1997.

Koerner, Michael (editor): Speech Recognition: The Future Now, Upper Saddle River, NJ: Prentice Hall PTR, 1997.

Lieberman, Protopapas, and Kanki, "Speech production and cognitive deficits on Mt. Everest", *Aviation, Space, and Environmental Medicine*, Vol. 66, No. 9, September 1995.

Mulder, G. and L. J. M. Mulder., "Information processing and cardiovascular control." *Psychophysiology*, 1981, 18: pp. 392-405.

Rabiner, Lawrence and Juang, Biing-Hwang: Fundamentals of Speech Recognition, Englewood Cliffs, NJ: Prentice Hall PTR, 1993.

Scanlon, M. V., "Acoustic monitoring pad for combat casualty care," June 1996, *Army Science Conference Proceedings*, Norfolk, VA.

Summers, W. V., "Effects of stress and final-consonant voicing on vowel production: articulatory and acoustic analyses," *J. Acoust. Soc. Am*., Sep. 1987; 82(3): pp. 847-863.