Sub-Auditory Speech Recognition

Kim Binsted
ICS Department, University of Hawaii

Charles Jorgensen
Code IC, NASA Ames Research Center

Abstract

Sub-auditory speech recognition using electromyogram (EMG) sensors is potentially useful for interfaces in noisy environments, for discreet or secure communications, and for users with speech related disabilities. Past research has shown that a scaled conjugate gradient neural network, using dual tree wavelets for feature transformation, can categorize EMG signals for small sets of individual words. Here we describe an attempt to recognize the individual phonemes of the English language. Recognition rates are significantly higher than chance; however, they are not high enough to be useful without further improvements, perhaps from using multiple neural networks on specialized recognition tasks, better sensor placement, and/or moving on to di/triphone recognition as is commonly used in audible speech processing.

Introduction

Speech recognition research and technology to date (e.g. Huang *et al*, 2001) has, quite naturally, focused on audible speech. However, there are many applications that call for sub-auditory speech recognition. In particular, sub-auditory speech recognition would be valuable in noisy environments (e.g. crowded rooms, cockpits), in environments where sound does not carry well or at all (e.g. underwater, near-vacuum), when discreet or secure communications are necessary or desirable (e.g. military applications, off-line comments during meetings), and for users with speech-related disabilities (e.g. vocal cord damage).

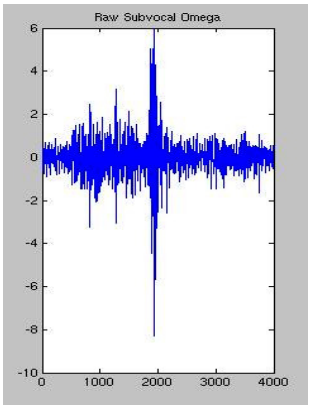

Figure 1: Raw EMG signal for subvocalized word "omega".

Relevance

NASA is interested in multi-modal interfaces as a way of increasing communication robustness and reducing information overload in human-human and human-agent systems. Imagine an astronaut exploring the surface of Mars, in collaboration with other human astronauts, a variety of robotic rovers, intelligent shipboard systems, and a mixed human-agent team back on Earth. In a space suit, both inputs and outputs are severely limited, and audible speech is by far the most convenient communications channel. However, audible speech interfaces cannot support many parallel interactions and, in the case of human-agent communication, are not very robust in the presence of noise, speaker stress, changes in gas mixture, etc. By adding sub-vocal speech to the repertoire of space suit interface designers, we hope to increase the robustness of audible speech by providing redundancy, and also to provide an alternate means of communication when appropriate (e.g. discreet communications) or necessary (e.g. a physiological problem renders the audible speech interface unusable).

Related Work

Little work appears to have been done on the usefulness of EMG sensors alone in speech recognition. Auditory speech recognition augmentation with EMG sensors, along the lines of that in our word experiments, was performed by Chan (2001), who proposed supplementing voiced

speech with EMG in the context of aircraft pilot communication. Chan studied the feasibility of augmenting auditory speech information with EMG signals recorded from primary facial muscles using sensors imbedded in a pilot oxygen mask. Chan used five surface signal sites during vocalized pronunciation of the digits zero to nine using Ag-AgCl button electrodes and an additional acoustic channel to segment the signals. This work demonstrated the potential of using information from multi-source aggregated surface sensors to improve performance of a conventional speech recognition engine.

Table 1: English language phonemes.

Phonemes			
Vowels	**Words**	**Consonants**	**Words**
ax	ago	*b*	big
ay	bite	*ch*	chin
uh	book	*k*	cut
aa	car	*d*	dig
ah	cut	*f*	fork
ey	day	*zh*	genre
ao	dog	*g*	gut
iy	feel	*hh*	help
aw	foul	*jh*	joy
ae	gas	*l*	lid
ow	go	*m*	mat
ih	hit	*n*	no
axr	percent	*p*	put
eh	pet	*r*	red
ix	sick	*sh*	she
uw	tool	*sh*	sit
oy	toy	*t*	talk
er	turn	*dh*	then
		th	thin
		v	vat
		w	with
		y	yacht
		z	zap

Approach

Here, we attempt to apply lessons learned from signal processing in general, and speech recognition in particular, to the problem of sub-auditory speech recognition. Using EMG sensors placed strategically in the throat area (see Figure 2), we hope to detect pre-speech EMG signals, and

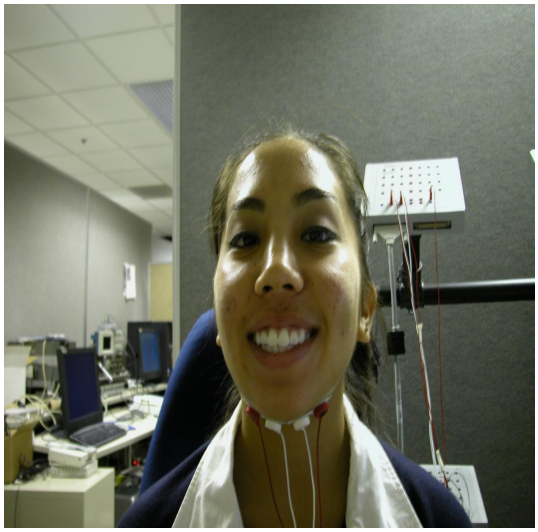

Figure 2: EMG sensor placement.

recognize the intended speech, even when it is not produced audibly. A key research question is: To what extent can audible speech recognition technologies and principles be applied to the problem of sub-auditory speech recognition?

Our past work (Jorgensen *et al*, 2003) has shown considerable promise. Using dual tree wavelets for feature transformation and a scaled conjugate gradient net for categorization, Jorgensen *et al* achieved 92% word recognition, over a set of six control words ("stop", "go", "left", "right", "alpha" and "omega"). A raw EMG signal from this experiment is shown in Figure 1. Similar results have been achieved for the digits ("zero" through "nine"). However, for a practical speech recognition interface, it is probably necessary to recognize sub-word patterns, because of the large number of words available to speakers.

Here, we establish a baseline for sub-auditory speech recognition by applying the approach described above to the task of recognizing *all* the phonemes used in the English language (see Table 1).

Data Collection

Two subjects, both female, were used. Two symmetrical EMG sensors (two electrodes each) were placed on the throat of each subject. Each subject was asked to sub-vocalize each phoneme, while thinking of the target word for that phoneme. For example, while sub-vocalizing the phoneme ***ao***, the subject would focus on the central vowel sound of the word "dog". On each subvocalization, the subject pressed a key, to record signal timing.

Table 2: Consonant recognition. The right "total" column shows the number of samples per phoneme in the test set, and the bottom "total" row shows the number of samples categorized as each phoneme. The yellow diagonal shows correct categorizations, and the red cells show errors over 20%.

	big	cut	fork	genre	gut	help	lid	mat	no	put	red	she	then	thin	vat	with	yacht		TOTAL
big	28	5	14	0	0	0	7	12	0	30	0	0	0	0	0	5	0		101
cut	3	44	0	0	15	0	9	0	3	0	15	3	0	0	3	6	0		101
fork	2	4	44	0	2	0	2	2	4	4	4	4	0	0	25	5	0		102
genre	0	4	0	38	0	0	9	0	18	0	13	11	0	0	0	7	0		100
gut	0	18	0	4	41	4	20	0	2	0	0	2	6	0	0	2	2		101
help	0	0	8	0	3	56	10	0	5	3	8	5	0	0	3	0	0		101
lid	0	7	2	0	5	0	74	0	0	7	2	0	0	0	0	0	2		99
mat	13	0	0	0	0	0	0	61	0	6	6	0	2	0	11	2	0		101
no	0	3	3	0	3	0	5	0	81	0	0	0	0	0	0	5	0		100
put	14	5	5	0	2	2	9	2	5	44	7	0	0	2	2	0	0		99
red	0	2	2	2	0	2	2	2	2	0	66	0	0	9	2	7	0		98
she	0	4	0	13	0	0	4	0	4	0	29	36	0	2	2	4	0		98
then	3	9	0	0	0	0	0	0	0	6	9	3	49	23	0	0	0		102
thin	0	0	0	0	0	0	0	2	0	0	2	0	14	79	2	0	0		99
vat	5	0	19	2	0	2	0	2	0	9	2	2	0	0	53	2	0		98
with	0	2	0	2	0	0	6	6	2	2	27	0	0	0	6	46	0		99
yacht	0	14	0	18	0	0	16	2	4	6	27	0	4	4	0	0	4		99
TOTAL	68	121	97	79	71	66	173	91	130	117	217	66	75	119	109	91	8		

Although vowel phonemes are syllables in their own right, consonant phonemes are difficult to pronounce (even sub-vocally) on their own. For this reason, subjects were asked to pronounce the consonant phonemes plus the vowel phoneme ***ax***. So, the consonant phoneme ***d*** would be pronounced as ***dax***, or "duh".

Each subject generated ten sets for each phoneme, each set consisting of approximately twelve examples. Since each of the two signal channels is processed independently, this leads to 12x10x2=240 signals per phoneme per subject.

Data Processing

From the raw data, 1.5 second-long samples were automatically extracted whenever the signal exceeded a threshold. Then, each sample was reviewed both automatically and manually, so that poor data (e.g. no subject key press indicated, or overlapping signals) could be removed. In order to introduce some temporal noise (which seems to help with training the network) into the sample set, we then shifted each sample a small amount right and left, resulting in five samples for each original

Table 3: Vowel recognition, labeled as in Table 2.

	ago	bite	book	car	cut	day	dog	feel	foul	gas	go	hit	percen	pet	sick	tool	toy	turn		TOTAL
ago	48	0	0	3	5	0	5	3	0	13	0	0	0	10	15	0	0	0		102
bite	6	27	0	0	0	6	0	27	0	3	0	0	0	3	3	0	18	6		99
book	28	5	26	3	0	0	0	0	0	3	5	3	3	0	3	10	10	3		102
car	12	5	2	7	5	14	7	0	0	21	0	7	2	7	7	0	0	5		101
cut	21	4	10	2	10	0	10	2	2	10	2	0	0	17	10	0	2	0		102
day	2	2	0	0	0	20	2	42	0	11	0	0	0	7	4	0	9	0		99
dog	12	5	0	7	0	0	32	5	0	17	2	5	0	7	5	2	0	0		99
feel	0	0	0	0	0	5	0	93	0	2	0	0	0	0	0	0	0	0		100
foul	0	2	2	2	2	0	4	10	56	4	6	0	0	2	0	8	0	2		100
gas	5	0	0	0	3	5	3	5	0	73	0	0	0	5	0	0	0	0		99
go	3	3	0	6	8	3	0	6	28	3	17	3	0	3	0	17	0	3		103
hit	13	3	3	0	0	3	0	8	0	18	0	5	3	23	20	3	3	0		105
percen	7	10	2	0	2	10	2	7	2	12	2	0	15	10	0	0	12	5		98
pet	9	2	0	0	0	0	12	2	0	19	0	5	2	28	21	0	0	0		100
sick	18	3	0	0	0	3	8	0	0	13	0	3	0	18	33	3	3	0		105
tool	3	9	3	0	0	0	0	0	9	6	15	3	3	0	0	29	15	6		101
toy	0	2	2	0	0	0	0	2	0	0	0	0	0	0	2	2	91	0		101
turn	2	7	0	2	2	2	0	24	7	5	2	0	2	2	0	5	7	29		98
TOTAL	189	89	50	32	37	71	85	236	104	233	51	34	30	142	123	79	170	59		

Table 4: Results of using specialized networks to distinguish between phonemic features. "Sampling problem" in the notes column indicates that at least one of the categories has too few samples for the categorization to be reliable.

Distinction	Categories	Success Rate	Epochs	Note
consonants vs. vowels	2	83%	2000	
CONSONANTS				
between alveolar consonants	7	75%	1000	
between plosive consonants	6	77%	500	
between fricative consonants	9	64%	1000	voiced vs. unvoiced are confused
between places of articulation	6	35%	500	probable sampling problem
between manners	7	n/a	n/a	sampling problem
pairwise, voiced vs unvoiced	2	100%	500	possible sampling problem
VOWELS				
between all simple vowels	10	63%	1000	
simple vs dipthong vs schwas	3	n/a	n/a	sampling problem
high+/low+/neither	3	83%	1000	
front+/back+/neither	3	83%	1000	
round +/-	2	81%	1000	errors are all false negatives
tense +/-	2	70%	1000	errors are all false positives
pairwise, simple vs. schwa	2	100%	500	possible sampling problem

sample, Then, a dual tree wavelet transform was applied for feature extraction. Finally, a scaled conjugate gradient neural network was trained (on 80% of the data) and tested (on 20% of the data).

Results

Initially, we used a single neural network to categorize all 40 phonemes for one subject. The resulting success rate was 18% (2500 iterations) - considerably better than chance, but not good enough to be useful for speech recognition. Then, we tried to categorize only the vowel phonemes.

Results from vocalized speech recognition led us to expect very poor results for the 18 vowels, and we were pleased to achieve a 36% success rate, using 2500 iterations in training (see Table 3). Not surprisingly, those pairs that human listeners find difficult to discern (e.g. ***ix***, "sick", vs. ***ih***, "hit"), were also often confused by our system.

The 22 consonants had a higher relative success rate of 33% after 1500 iterations (see Table 2). It seems that *voicing* is not easily detected by the system, in that confusable pairs (e.g. ***d***, voiced alveolar plosive, and ***t***, voiceless alveolar plosive) are often distinguished only by voicing. Also, the network seems to put difficult-to-identify items into "garbage can" categories, and seemed to have particular difficulty with the alveolar consonants.

The above results suggested that the phonological features used to distinguish vocalized phonemes might also be relevant for subvocalized phonemes, and that a productive approach might be to train specialized networks to categorize the phonemes based on these features. These networks could then be arranged into a decision tree, which could then be used to categorize all 40 phonemes effectively.

The next step, then, was to see if specialized networks could recognize the key phonological features. We have not yet completed this step, but partial results are given in Table 4. In several cases, one or more categories had very few samples, leading to unreliable results. This problem can only be corrected by taking more samples.

These partial results suggest that a decision tree something like the one in Figure 3 might be effective at categorizing the full phoneme set. If the accuracy estimates at each node are correct, then this tree would have an overall accuracy of 50-60%. Note that both the binary features for the vowels and the combination of manner and place of articulation for consonants are quite redundant, so that it should be possible to lower error rates by looking for feature combinations.

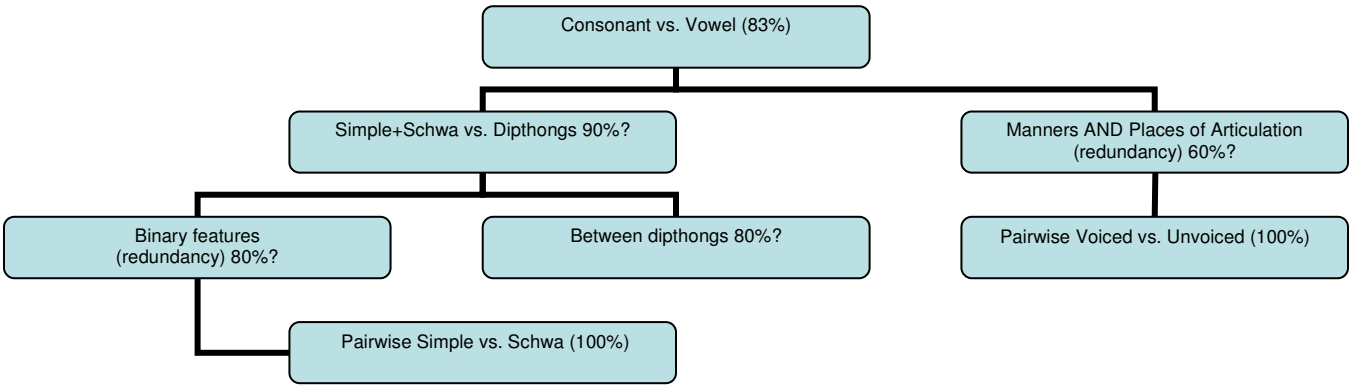

Figure 3: Possible decision tree, using specialized networks that categorize based on phonemic features. Question marks indicate hoped-for accuracy once sampling problem is solved.

Further Work

The immediate next step is to complete the above analysis, and determine whether or not the decision tree approach will reduce error rates. Also, we should experiment with sensor placement – perhaps some problem feature distinctions can be eliminated or reduced if the sensors are placed to detect key muscle movements.

We have, however, achieved our primary goal, which was to establish a baseline for sub-auditory speech recognition. Also, it seems that features of spoken speech that are relevant to auditory speech recognition are also relevant to sub-auditory speech recognition (i.e. using an EMG signal). This suggests that techniques which have proven useful in processing spoken speech, such as diphone or triphone recognition, would also be useful in processing sub-auditory speech. We plan to explore further in this direction.

Acknowledgements

This material is based upon work supported by the National Aeronautics and Space Administration through the NASA Astrobiology Institute under Cooperative Agreement No. NNA04CC08A issued through the Office of Space Science. It was also supported in part by the NASA Faculty Fellowship Program.

Bibliography

A. Chan, K. Englehart, B. Hudgins, and D.F. Lovely (2001). "Myoelectric Signals to Augment Speech Recognition," *Medical and Biological Engineering & Computing*, pp. 500-506 vol *39(4).*

Huang, X., Acero, A., and Hon, H-W. (2001). *Spoken Language Processing: A guide to theory, algorithm and system development*. New Jersey, Prentice-Hall PTR.

Jorgensen, C., Lee, D., and Agabon, S. (2003). "Sub Auditory Speech Recognition Based on EMG/EPG Signals." In Proceedings of International Joint Conference on Neural Networks, Portland Oregon, July 2003.