Phone Recognition for Mixed Speech Signals: Comparison of Human Auditory Cortex and Machine Performance

Shuo-Yiin Chang[§*], Erik Edwards[*ǂ], Nelson Morgan[§*], Dan Ellis[*ɣ], Nima Mesgarani[ɣ], and Edward Chang[ǂ]

TR-15-002

June 2015

Abstract

It is well known that human beings can often attend to a single sound source within a mixed signal from multiple sources, and that unaided automatic speech recognition (without the benefit of effective blind source separation) is quite poor at this task. Here we report on the analysis of human cortical signals to demonstrate the relative robustness of these signals to the mixed signal phenomenon, which is contrasted to a deep neural network-based ASR system. Confirming this difference with a carefully designed experiment is the first step towards ultimately improving blind source separation for the purpose of speech recognition; in particular, the design of features extracted from the neural signals is leading to insights about the corresponding feature extraction on the acoustic side, e.g., for CASA systems of the future.

§ EECS Department, UC Berkeley, Berkeley, CA, USA
* International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, California, 94704
ǂ Department of Neurological Surgery, UC San Francisco, San Francisco, CA, USA
ɣ Electrical Engineering Department, Columbia University NY, NY, USA

This work was partially supported by funding provided to ICSI through National Science Foundation grant IIS: 1320260 (“Towards Modeling Source Separation from Measured Cortical Responses”). Additional funding was provided to UC San Francisco through National Science Foundation grant IIS: 1320366 (“Towards Modeling Source Separation from Measured Cortical Responses”). E.F.C. was supported by National Institutes of Health grant R01-DC012379, and McKnight Foundation. Edward Chang is a New York Stem Cell Foundation - Robertson Investigator. This research was supported by The New York Stem Cell Foundation. We also thank Liberty Hamilton and Zack Greenberg for their assistance with the ECoG work. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors or originators and do not necessarily reflect the views of the National Science Foundation, the National Institutes of Health, or the New York Stem Cell Foundation.

1. Introduction

We have recently been studying the cortical mechanisms involved in auditory source separation for mixed single-channel speech signals, using neuroelectric responses directly measured from the surface of the human cortex using a 256-electrode array, giving a view of cortical sound processing of unprecedented detail and flexibility. This work has dual goals: (1) to better understand these neural mechanisms in humans, and (2) to use this new understanding to help us to design better artificial systems for the separation of mixed signals (e.g., to voices), with the ultimate goal of making speech recognition systems more robust. In both cases, the effort has built on the earlier work reported in [1], in which the spectrotemporal representation of attended speech was reconstructed. In the newer work, our goal is to design a CASA (computational auditory scene analysis) system with insights from the analysis of the human neural data. For this document, however, we are reporting an initial result in which we have observed that signals collected from the surface of the human cortex (specifically in the area called the Superior Temporal Gyrus, or STG) can be processed to classify phones with similar error rates for both single and mixed single cases. The rest of this paper describes the experiment, interprets the results, and discusses implications for future work.

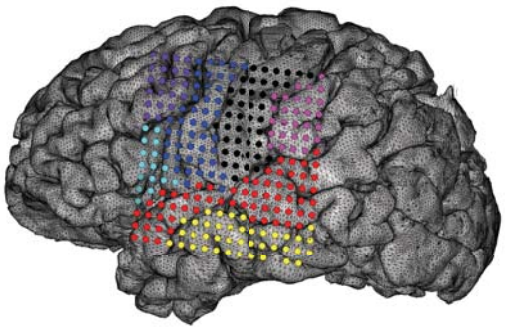

Figure 1: *MRI reconstruction of subject's cortex, with electrocorticogram (ECoG) electrodes (16x16 grid, 4 mm spacing) indicated by dots. The red and yellow regions (temporal lobe) are expected to contribute most to speech processing.*

2. Experimental Methods

2.1. Data

The acoustic and neural data used in this study were previously generated in sessions with surgery patients as described in [1]. The patients had customized high-density electrode arrays implanted subdurally that are sometimes necessary for the surgical management of patients with epilepsy refractory to medications [2] (see Figure 1). They participated in behavioral testing using stimuli from the Coordinate Response Measure (CRM) corpus [3]. The corpus consists of phrases of the form "Ready (call sign) go to (color) (number) now" spoken with different combinations of call signs ("Tiger" or "Ringo"), 3 colors ("Blue", "Green", "Red"), 3 numbers ("Two", "Five", Seven") and two speakers. The patients are instructed to report the color and number associated with a call sign (e.g. "Tiger"); however, they do not know *a priori* which speaker will be the target in each trial. Consequently, subjects are required to attend to both speakers at the beginning of mixture until they hear the target call sign and then attend to the corresponding speaker. At the end of each experiment block, the patients have responded to the same 28 sound mixtures while attending to both speakers. This experimental design allows us to determine the effect of attention on neural responses, while controlling for identical acoustic stimulus conditions (i.e., hearing the same speaker mixture, while attending to only one or the other voice). For the purposes of this study, we have been working with data collected from three subjects, but since neural responses are highly individual, all the results reported here are from a single individual.

Given the extremely modest amount of data (e.g., 3115 phones) from the CRM experiments, we also used of a subset of TIMIT to augment the acoustic training set. As described below, we also used a jackknifing technique [4] to make better use of the limited amount of data.

2.2. Feature extraction

*2.2.1 Acoustic feature*s

For the purposes of this study, we only used un-enhanced MFCCs. We used the Kaldi front-end [5] to produce a 39 dimensional feature vector every 10 ms, which converted each 25 ms signal frame into 13 Mel-cepstral coefficients, including energy, plus their first and second differences.

*2.2.2 Neural feature*s

While the usual signal representations for phone or speech recognition are *acoustic features*, we also attempt to decode the phone sequence directly from *neural features*. We use the term *feature* here in the pattern recognition sense of any suitably preprocessed input to the recognizer.

The question is how to suitably preprocess the raw, often noisy, broadband ECoG signal for ~optimal performance in the recognizer. First, some degree of noise is unavoidable in the present clinical context, and some epochs and electrodes are rejected after human examination [as in 1]. Second, although human brain waves have been known since 1929, and intraoperative recording attempted soon thereafter [6], it was not until relatively recently [7, 8] that it became apparent that the information available in the so-called "high-gamma" band (~70-170 Hz) carries information of far greater spatial and temporal specificity compared to the more traditionally-studied frequency ranges below ~60 Hz [9]. Although we have not ruled out the use of lower frequencies as auxiliary information [e.g., 10], the present study uses only the ECoG data in the high-gamma range. Specifically, we sum over Hilbert envelopes [11] for center frequencies ~70-170 Hz [as in 9, 12], and downsample to 100 Hz. Analytic amplitudes are distributed ~Rayleigh and heteroscedastic [13, 14], so we further take the natural log to render ~Normal and to stabilize variance [15, 16].

The end result of preprocessing is a set of ~250 high-gamma time series, one for each retained electrode. With small training sets, DNNs may be prone to overfitting, so dimensionality reduction to 48 neural features was deemed necessary. We then explored several approaches to reduce 250 single-electrode features to 48 neural features.

The most obvious, and nearly the oldest, method of dimensionality reduction is spatial principal component analysis (PCA). However, the resulting components lack any physiological significance and also serve poorly in a pattern recognition sense. This problem has been known for nearly as long as spatial PCA itself [17, 18], and the most frequent solution has been to *rotate* the components so as to achieve sparsity, locality, clustering by similarity, or some other objective. After extensive study of available rotation methods, component analyses, hard and soft clustering approaches, and various other machine learning methods surrounding embedding and dimensionality reduction, it was surprising to find that the old method of *varimax rotation* [19, 20] performed as well or better than any of the several modern methods tried. This becomes less surprising when one considers that such rotation is the old L2-norm method to achieve sparsity (called "simple structure" in factor analysis), co-sparsity [21], within-cluster smoothness, across-cluster decorrelation, and greater physical plausibility; all of these known to be virtues for pattern recognition features [e.g., 22]. As an L2 method, it is also extremely fast.

Convex non-negative matrix factorization (NMF) [23] was found to slightly improve the results taking the varimax components as initializing input. Neither ordinary NMF [24], nor convex NMF without a good initialization, performed as well as varimax for our purposes. The requirement of a good initialization is the known drawback of NMF. Intuitively, convex NMF achieves the clustering objective that correlated electrodes should cluster together, whereas the resulting cluster time series should be as decorrelated as possible.

The final neural features used here are thus 48 convex NMF components derived from varimax rotation of spatial PCA analysis of ~250 log-high-gamma time series (see Figure 2).

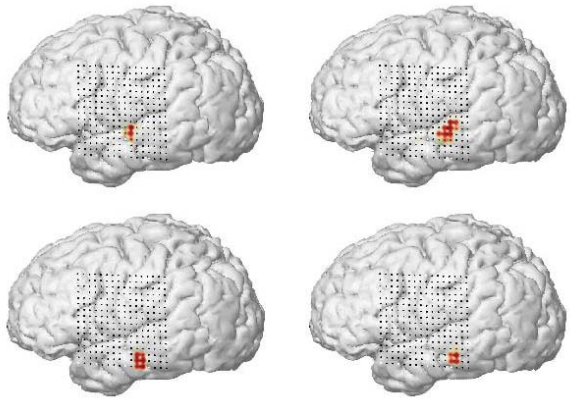

Figure 2: *Spatial weightings for 4 out of 48 of the convex NMF neural features used. These are typical examples, chosen for their loadings onto temporal lobe sites important for speech processing*.

2.3. Phone recognition system

For both neural and acoustic observations, we exploited the hybrid HMM/Artificial Neural Network (ANN) architecture [25] (more recently called HMM/DNN [26,27] when more than a single hidden layer is used) to model context-independent phoneme as shown in Figure 3 where network outputs were used as posteriors to derive emission probabilities for hidden Markov models (HMMs). We used the Kaldi toolkit [5] for both model training and decoding, as well as for the ANN processing. The hybrid HMM/ANN set up was adapted from Kaldi recipe s5 [5].

The inputs of the ANNs were obtained from splicing 39-d MFCCs or 48-d neural features across 17 frames, followed by reducing the dimension to 250 using linear discriminant analysis. Mean and variance normalization were performed for both MFCCs and neural features. The ANNs had 2 hidden layers (marginally "deep"), each of which consists of 1100 *tanh* units. The output layer consisted of 117 context-independent phonetic states (three states per phoneme), giving 1.6 M parameters in total (see Figure 3). Frame-level forced alignment was provided by a simple context-independent HMM/GMM system. The ANNs were trained with stochastic gradient descent, starting with a learning rate of 0.015 and ending at 0.002. During training we decrease them by a factor of 1.14, except for 5 epochs at the end during which we kept them constant. The network was trained for a total of 20 epochs. A biphone language model was estimated on the training set.

The acoustic recognition scenario gave a 26.9% phone error rate for the standard TIMIT train-test set, where 3696 utterances were included for training and 192 utterances for testing. Given the constraint of ECoG neural data, we set up our train-test sets using only 374 utterances from TIMIT, 155 utterances from single source CRM data and 61 utterances from mixed source CRM data.

Due to the small amount of data, we used the jackknife resampling process [4], splitting the set of CRM data into 5 different train-test cuts. For each of the training sets, we randomly drew samples from the CRM data and augmented them with 374 utterances of the TIMIT set. The rest of the CRM sets were used for testing. This yielded 7,520 instances of phones for single source test sets and 3,115 instances for mixed source test sets. The results reported here are average score over the 5 different train-test sets.

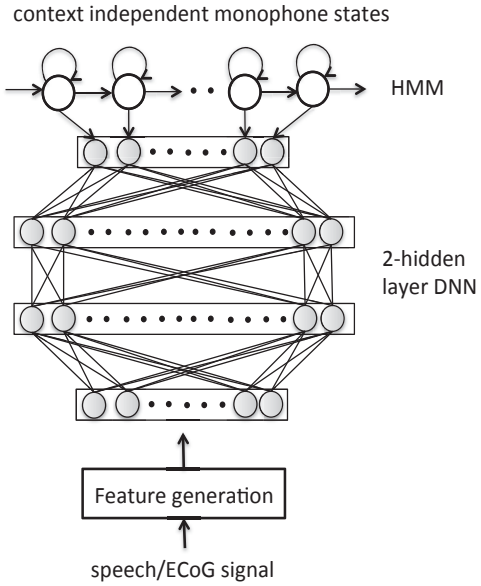

Figure 3: *Hybrid HMM/ANN architecture for context independent monophone model where acoustic or neural features were used as input of 2-hidden layer deep neural network to model 117 state targets*

3. Results and Discussion

Table 1 below shows that the phone error rate for the acoustic features, while far better than that achieved with our neural signals for the single voice case, degrades greatly for the mixed signal. For the neural features, the error rates overall are quite high, but there is very little additional degradation when the subject is motivated to attend to the desired voice. Why should this result be interesting, when it conforms to our expectations that humans will tend to be more robust to interfering signals than our artificial systems? Let's consider what we were trying to see.

The primary hypothesis being tested was, given specific neural signals, and given the chosen features extracted from these signals, that the phone recognition error rates would be affected far less for the neural signals than for the acoustic signals. While it is a common experience that humans do better in the "cocktail party" scenario than our current machine methods, what is being tested here is the utility of the specific brain signals that we are measuring to show this phenomenon. Furthermore, the experimentation with feature extraction methods given the raw data has begun to show us what aspects of the STG neural signals are most effective for phone recognition.

	Acoustic features	Neural features
Single source	48.6%	68.2%
Mixed source	73.2%	70.5%

Table 1: *Phone error rate for complete CRM utterances*

We currently make no use of standard enhancement or blind source separation algorithms. Furthermore, for consistency with the small amount of neural data we could collect, the task-specific acoustic data set was quite small. Consequently all error rates are quite high. Nonetheless, the observed effects are quite striking: the neural features yield nearly the same error rates for single and mixed sources, and the acoustic features are much less informative for the mixed source case.

Tables 2 and 3 show similar trends for the phone error rates in target words (color and number, respectively). Phone error rates for acoustic features rise significantly for the mixed signal case, but are nearly the same using neural features.

	Acoustic features	Neural features
Single source	54.0%	72.0%
Mixed source	75.2%	73.1%

Table 2: *Phone error rates within the color target word (red, green, or blue)*

	Acoustic features	Neural features
Single source	56.1%	73.5%
Mixed source	79.9%	73.9%

Table 3: *Phone error rates within the number target word (two, five, or seven)*

4. Where this work can lead

The work reported here confirmed that the STG can provide meaningful information for phonetic recognition, at least for the task used, that can be relatively independent of interfering voices that the subject is not paying attention to. This is in contrast to the oft-observed increases in error rate for phone recognition given such interfering signals.

Despite this, it is likely that the phone categories used here are not the best match to the observed neural signals. Other work [28] has shown that the instrumented areas appear to provide information about, for instance, manner categories. We also are not yet certain that we are computing the best neural features; we are currently focusing on the "high gamma" region of the neural signal (~70-170 Hz), and there may yet be some utility in observing signals in other bands, something that we are starting to explore.

Ultimately, our goal is to learn more about how the brain is recognizing speech when competing signals (particularly multiple voices) are present. But as our other goal is to learn from this exploration how we can improve automatic speech recognition in the presence of mixed signals, we will have to use partial information (for instance from comparison of error patterns) to modify our artificial systems to act more like the natural one. We are just beginning this process, and this paper reports the initial effort to simultaneously study the scientific and engineering components.

5. References

[1] Mesgarani N, Chang E (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485: 233-236.

[2] Chang E, Rieger J, Johnson K, Berger M, Barbaro N, Knight, R (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience.*

[3] Bolia RS, Nelson WT, Ericson MA, Simpson BD (2000). A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America* **107**, 1065.

[4] Efron B (1982). The jackknife, the bootstrap, and other resampling plans. *Society of Industrial and Applied Mathematics CBMS-NSF Monographs*, 38.

[5] Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlícek P, Qian P, Schwarz P, Silovsk J, Stemmer G,

Vesey K (2011). The kaldi speech recognition toolkit. *Proc. IEEE 2011 Workshop on ASRU*. Dec. 2011, IEEE Signal Processing Society.
[6] Berger H (1931). Über das Elektrenkephalogramm des Menschen. Dritte Mitteilung. *Arch Psychiatr Nervenkr* 94: 16-60.
[7] Crone NE, Miglioretti DL, Gordon B, Lesser RP (1998). Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. *Brain* 121(Pt 12): 2301-15.
[8] Crone NE, Boatman D, Gordon B, Hao L (2001). Induced electrocorticographic gamma activity during auditory perception. *Clin Neurophysiol* 112(4): 565-82.
[9] Edwards E (2007). Electrocortical activation and human brain mapping. Dept. of Psychology dissertation. Berkeley: Univ. of California: 147 p.
[10] Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron* 77(5): 980-91.
[11] Baghdady EJ (1961). Diversity techniques. In: Baghdady EJ, ed. Lectures on communication system theory. New York: McGraw-Hill**:** 125-75.
[12] Edwards E, Soltani M, Kim W, Dalal SS, Nagarajan SS, Berger MS, Knight RT (2009). Comparison of time-frequency responses and the event-related potential to auditory speech stimuli in human cortex. *J Neurophysiol* 102(1): 377-86.
[13] Arens R (1957). Complex processes for envelopes of normal noise. *IRE Trans Inf Theory* 3(3): 204-7.
[14] Bendat JS, Piersol AG (2000). Random data: analysis and measurement procedures, 3rd Ed. New York: J. Wiley.
[15] Prucnal PR, Goldstein EL (1987). Exact variance-stabilizing transformations for image-signal-dependent Rayleigh and other Weibull noise sources. *Appl Opt* 26(6): 1038-41.
[16] Pitas I, Venetsanopoulos AN (1990). Nonlinear digital filters: principles and applications. Boston: Kluwer Academic.
[17] Lorenz EN (1956). Empirical orthogonal functions and statistical weather prediction. Statistical Forecasting Project. Scientific Report No. 1. Cambridge, MA: MIT, Dept. of Meteorology: 49 p.
[18] Jolliffe IT (2002). Principal component analysis, 2nd Ed. New York: Springer.
[19] Kaiser HF (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23(3): 187-200.
[20] Richman MB (1986). Rotation of principal components. *J Climatol* 6(3): 293-335.
[21] Elad M (2012). Sparse and redundant representation modeling – what next? *IEEE Signal Process Lett* 19(12): 922-8.
[22] Hall MA (1999). Correlation-based feature selection for machine learning. Dept. of Computer Science. Hamilton, New Zealand: University of Waikato: 178 p.
[23] Ding C, Li T, Jordan MI (2010). Convex and semi-nonnegative matrix factorizations. *IEEE Trans Pattern Anal Mach Intell* 32(1): 45-55.
[24] Lee DD, Seung HS (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755): 788-91.
[25] Bourlard H, Morgan N (1993). Connectionist Speech Recognition: A Hybrid Approach, Kluwer Press.
[26] Mohamed A, Dahl G, Hinton G (2012). Acoustic modeling using deep belief networks. *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 14 –22.
[27] Seide F, Li G, Yu D (2011). Conversational Speech Transcription Using Context-Dependent Deep Neural Networks. *Proc. Interspeech*, 2011.
[28] Mesgarani N, Cheung C, Johnson K, Chang EF (2014). Phonetic feature encoding in human superior temporal gyrus. *Science* 343(6174): 1006-10.