C. IDENTIFICATION AND SIGNIFICANCE OF THE PROBLEM OR OPPORTUNITY

Speech communication is so valuable to humans because it is possible in a very wide range of circumstances, including all kinds of adverse acoustic conditions including unrelated background noises, highly reverberant environments, and competing incidental speech. For speech interfaces to machines to reach their full potential, the machines must be able to understand spoken commands in a range of situations comparable to (and ideally wider than) those tolerable by human listeners.

With the appropriate interface design, speech recognition can provide greatly enhanced usability and efficiency, but a significant level of recognition errors, arising from, for instance, interfering speech or noise will rapidly erase this advantage.

Progress in automatic speech recognition (ASR) technology has been achieved by simplifying and constraining the task until meaningful progress could be made, then gradually rolling back these artificial restrictions as techniques (and computational power) have improved. The earliest speech recognizers addressed the problems of discriminating just a few words, carefully spoken by known speakers, in ideal acoustic environments (head-mounted microphones and silent backgrounds). Our recent progress has removed many of these restrictions: Current state-of-the-art systems can recognize tens or hundreds of thousands of words spoken by arbitrary voices in natural, conversational style. The constraint of idealized acoustics, however, has proven problematic to relax. Amazingly, current speech recognition systems show significant degradation at signal-to-noise ratios (SNRs) of 20 dB, and even the best noise compensation schemes leave ASR systems next to unusable at 10 dB SNR – a noise level which would hardly distract a human.

One reason that noisy backgrounds prove so challenging, and why robust speech recognition in noisy environments seems so elusive, is that the “ideal acoustics” constraint led to a range of techniques that relied on the assumption that the entire acoustic signal was informative about the particular target speech sound being uttered. When noise is introduced, some or all of the signal observed in any time window may be completely irrelevant, since it relates to an unconnected source.

We believe this problem can only be completely overcome by an approach to recognition that intrinsically recognizes the partial availability of relevant information in the signal, as we will describe in detail in this proposal. Washington Software, Inc. (WSI) and Columbia University proposes an analytical/experiment approach in this Phase I STTR program to evaluate these techniques in combination with other, existing approaches to deal with noisy speech.

D. PHASE I TECHNICAL OBJECTIVES

The essence of speech recognition is to compare a description of the observed signal with previously-learned models corresponding to particular speech sounds. The problem of speech recognition in chaotic acoustic environments is that the observed speech signals no longer resemble the models as a result of the environmental corruption, which takes the forms of both independent additive noise (noise from machinery, weather, and other voices), and distortion of the voice signal resulting from multi-path reflections (i.e., reverberation).

There are a few different kinds of approach that can be taken to address this problem:
• Signal enhancement: Improve the estimation of the original clean speech signal, either by using multiple microphone signals to take advantage of the compact spatial origin of the source, and/or by using a model of the intrinsic structure of the signal (such as its periodicity) to remove interference energy.

Signal enhancement can be applied in several domains, such as the raw waveform level, or the speech feature vectors; the particular principles being exploited may be more or less difficult to apply in different domains. Multi-channel waveform-level enhancement includes acoustic beamforming (ABF), several independent component analysis (ICA) approaches, and model-based inference techniques such as (Attias and Deng, 2002); single channel approaches include so-called computational auditory scene analysis (CASA) techniques, which most commonly rely on extracting energy related to a single underlying periodicity, and model-based time-frequency filtering such as (Roweis, 2003).

• Robust feature space: Model matching is performed in a particular feature space chosen to optimize computational and performance goals. Better feature representations provide the most phonetically-relevant information in the most compact feature space (to simplify model learning and matching), but alternative features may exhibit greater invariance to potential corruptions such as additive noise or channel coloration, at differing costs in terms of dimensionality or relevant information content. Examples of noise-robust features include mean-normalized Mel-frequency cepstral coefficients (MFCCs), perceptual-linear-prediction (PLP) cepstra, and modulation-filtered spectrogram features (MSG).

• Model adaptation: Adjust the reference models in the recognizer to be a closer match to the observed signals. This is typically done by estimating a few parameters that apply across the whole set of models, so that after adjusting the model set to match the gross properties of the observations, the discrimination between individual model components still accurately classifies the speech sounds. Both maximum-likelihood linear regression (MLLR) and parallel model combination and its relatives (PMC) fall under this description.

Another approach to improving the match between models and noisy observations is simply to use examples of the noisy signals in training the original models. When the range of noisy conditions can be accurately anticipated, and when the noise properties are relatively stationary, this approach can be very effective as shown in the “multicondition training” track of the recent Aurora tasks. When these conditions do not hold, however, the exponential combinatorics of attempting to train models for each speech sound combined in all possible temporal alignments with all possible interference conditions rapidly become intractable.

• Model matching: The core of a statistical pattern recognizer is a function to calculate the likelihood of an observation given a hypothesis of the ‘true’ original signal state (model). Typically, this reduces to a Euclidean distance in a suitably transformed feature space, but the effect of additive corruptions can be directly accommodated in this stage with a likelihood calculation that adjusts its results based on whether a particular feature dimension is judged as dominated by the target voice or by some other interference. This is the key idea behind the “missing-data” recognition techniques.

A critical factor influencing the performance of any speech recognizer is the quality of the original sensor signals, typically obtained from one or more microphones. For a given environment, the closer a microphone is to the speech source – the speaker’s mouth – the greater the relative energy of the target signal compared to interference both from other acoustic sources and from reverberant reflections. However, in a given application it may be undesirable or even impossible to collect such close-talking microphone signals.
For this Phase I STTR project, we propose to evaluate a range of approaches drawn from the above options to find the best trade-off between practicality, generalizability, computational expense, and recognition performance. Specifically, our technical objectives in Phase I will be as follows:

- Collect and simulate data for a range of microphone/sensor configurations. Specifically, we will consider a head-mounted mic, a push-to-talk “stalk” mic, and a small array of mics built-in to display equipment (e.g. 4 mics arranged at the corners of a video screen). Our experiments will particularly consider how the array can be used in conjunction with a single close-talk mic.
- Implement several different signal enhancement schemes: ABF (using multiple microphones to suppress non-target energy), ICA (using multiple microphones to enhance a statistically independent source), and model-based enhancement (projecting the noisy signal onto a subspace defined by clean speech properties).
- Implement several different noise-robust feature schemes (normalized MFCCs, PLP features, MSG features, and spectral features orthogonalized by differencing in time and frequency).
- Implement several different model compensation schemes including MLLR and PMC.
- Develop missing-data recognition and multi-source decoding implementations to focus recognition on portions of the signal that are most informative about the target speech.

Of all our investigations, this final point is the most innovative, and, we feel, the most likely to yield dramatic improvements. Experiments have shown that when given oracle knowledge about which time-frequency regions are dominated by the target, missing-data recognizers can give results very little degraded from the clean-signal case. Obtaining good estimates of this available-data mask is a significant challenge. Here, we propose to combine spatial information from multiple microphones, with intrinsic signal structure constraints of the kind used in CASA systems, which are integrated with prior knowledge about the statistical structure of speech signals through the mechanism of the multi-source decoder. This approach bypasses the attempt to completely recover the original speech signal - which, in many adverse environments, is evidently doomed - and instead concentrates on using the available reliable information from the target speech to infer the words being spoken. This project is the first opportunity to combine all these sources of information and techniques in an approach to solving a real-world speech recognition problem.

E. PHASE I WORK PLAN

E.1 Introduction

Acoustic Test Data: In order for our investigations and comparisons to provide accurate information about the performance in the kinds of real-world applications being considered, we require test data that are representative of such chaotic acoustic environments. Accurate test results require large test sets, and speech recognition systems typically demand as much task-matched training data as can be obtained; existing noisy-speech databases (such as Aurora and SPINE) are limited in their size, relevance, and realism. Moreover, multiple acoustic sensors are almost certainly required for these demanding applications, and the few multi-channel speech corpora (such as the recent Meeting Recorder data made available by LDC) do not address high noise conditions.

Thus, an important foundation of our Phase I work will be the development of a sizeable corpus of speech signals in realistic chaotic acoustic environments that we can use to develop and test our various approaches. For maximum realism, these data would all be directly recorded in real-world conditions. However, recordings of this type on the scale required would be both expensive and inflexible, since properties such as the basic target-to-interference ratios cannot be modified once the recordings are made.
Instead, we propose to develop algorithms to simulate these signals with very high accuracy, at the same time as providing parametric control over the nature, number, and level of interfering sources, sensor microphones, etc. We will synthesize these signals using ‘virtual acoustics’ techniques, modeling the impulse response couplings between multiple virtual sources and sensors.

In order to ensure the realism of these simulations, we will base them on a small amount of actual field recordings. Recordings from small microphone arrays in a few real-world chaotic environments (the inside of a vehicle traveling in bad weather, or a room full of speakers such as a coffee shop) will provide us with realistically “diffuse” multi-microphone background noise, and actual reverberant impulse responses measured by recording pseudo-white-noise sources at different locations in quite conditions.

Combining all these factors will allow us to generate synthetic examples of arbitrary target speech as if spoken and recorded in these specific (and similar) real-world environments. These simulations will be verified by comparing a number of overall statistics, including speech recognizer performance, with a limited amount of real speech recorded in the original environments. Once the synthesis path is developed, however, we will be able to produce essentially unlimited amounts of highly-realistic multi-channel data for both test and training.

**Speech Evaluation Approach:** For the initial phase of this project we must choose a specific speech recognition task to evaluate, but this need not mirror the full real-world task that is our ultimate goal. By using a relatively small-vocabulary recognizer in our initial evaluations, we will demonstrate the effectiveness of our enhanced recognition techniques without incurring the large computational costs of fully general speech recognition. There is little in our set of techniques that has any substantial dependence on the scale of the speech recognition task, so we are confident that improvements demonstrated in a limited-vocabulary context will generalize to less constrained speech tasks.

We propose to conduct our initial experiments using continuous digit strings. The TIDIGITS corpus, originally recorded in 1984, has long been a mainstay of speech recognizer evaluation, and in 1999 it was chosen by the European Telecommunications Standards Institute as the basis for their AURORA noise-robust feature evaluations; although word error rates are below 1% for digit strings recorded in quiet, added noise is still such a difficult problem that a task as highly constrained as digits is needed to get error rates that are not already saturated (i.e. close to chance) for SNRs in the -6..10 dB range. By mirroring the AURORA task (albeit with more difficult noise backgrounds, and with the additional information from multiple sensors) we will also have a large body of existing results based on the best current noise compensation schemes from the 2001 and 2002 evaluation rounds that we can compare against.

One drawback of a very small vocabulary task such as digits is that its phonetic content is unbalanced and most sounds occur in only one or two contexts. Depending on progress with the digits task, we may go on to look at a more complex task such as the Wall Street Journal read-speech task (WSJ), which contains a rich selection of phonetic contexts. WSJ was the basis for another AURORA evaluation task, so, as with the AURORA-style digits, there is already available a “reference” configuration of a standard conventional speech recognition package (the public-domain MSU ISIP recognizer in the case of WSJ (Parihar and Picone, 2001)) that we can use to jump-start our recognition experiments.

**Distinguishing Spoken Commands:** The techniques described below address the problem of recognizing words spoken by a target voice in a chaotic environment that may include incidental speech. While this
capability is necessary for any voice-control system to be used in those conditions, it is not a complete solution in itself; in general, there may be several recognizable voices simultaneously present, and the speech interface must decide which, if any, of the utterances are intended as commands for the system. Approaches to making this decision include:

- Dominant voice: In the case of head-mounted or stalk microphones, the target voice is easily identified as the strongest signal present. If a push-to-talk button is included, the system-directed utterances are also trivially identified. Even in these cases, however, there will often be significant interference energy in the observed signal, so mixture-robust recognition techniques are still required. Directional filtering: Scenarios without a close-talking mic, that rely instead on an array of mics embedded in the equipment, can use spatial cues to define a point in space from which commands must originate, so that the user knows to speak from that location when addressing the system.

- Prosodic cues: More sophisticated discrimination between machine-directed and other utterances can be based on analysis of the verbal and nonverbal attributes of the identified target voice. System commands or requests will often exhibit stereotypical intonation and phrasing patterns, and operators may easily learn to address the machine with a distinctive “tone of voice”.

- Monitoring interactions: If we consider how human participants identify which utterances are directed to them, one factor is the conversational context: an utterance that is clearly part of an ongoing dialog between two other people should not, in general, be interpreted as a separate command. A high performance voice separation and recognition system might be able to track these non-system-directed utterances to improve its ability to distinguish commands.

- Dialog design: Although the above techniques provide technical approaches to identifying appropriate utterances, often the greatest impact can be gained from careful design of the user interface, such as the command phrases, dialog management, and nonspeech feedback used by the system. Simple techniques such as a “Computer, ...” prefix for command interactions (familiar to viewers of Star Trek) may provide the best trade-off between operator convenience and context disambiguation.

Although the issues of verbal command identification and user interface dialog design are critical to the success of an overall deployment, we view them as orthogonal to the problem of accurately recognizing the noise-corrupted speech, and we see this latter problem as a far greater challenge. Hence, the rest of this proposal, and the work in phase I of the project, is directed specifically at the specific issue of recognizing a target voice in a chaotic acoustic environment.

E.2 Detailed Technical Approach

In this section we present greater detail on the central elements of our approach to explain why it can significantly improve ASR in chaotic environments. Our approach relies on three key components: missing data recognition, to match the observed speech features to learned models only over the regions where the target is judged to be dominant; estimation of those regions based on spatial source location estimation and intrinsic signal characteristics such as periodicity; and further refinement of those regions by comparison with the existing speech models using the techniques of multi-source decoding. Each of these stages is described in more detail below:

**Missing Data Recognition:** Our approach starts with the idea of a set of models of the individual sounds that can occur in our mixtures, similar to the models used in speech recognition. The classical application of statistical pattern recognition is to find a maximum a-posteriori probability fit across a range of class models to a set of signal features X i.e. the features are interpreted as an instance of the model, where:
Rearranging via Bayes’ rule gives:

\[ M^* = \arg \max_{M_i} P(M_i | X) \]  

(1)

where the prior term \( P(X) \) does not vary across the models, so can be dropped. Individual classes \( M_i \) are thus represented by distribution models \( P(X|M_i) \), which are a convenient way to represent prior class knowledge: The feature values observed in a set of training instances are generalized, typically as Gaussian mixture models (GMMs). The implicit assumption is that observations at classification time will be fully and directly comparable to the training examples on which the distribution models are based.

In the case of sound mixtures, however, any single target sound may appear in an infinite variety of acoustic contexts, formed by different combinations of different background sounds. We may describe this mathematically by defining a new variable, \( Y \), as the actual feature observations of the total, compound mixture, and our classification problem becomes:

\[ M^* = \arg \max_{M_i} P(M_i | Y) = \arg \max_{M_i} P(Y | M_i) P(M_i) \]  

(3)

One approach to recognizing the target sound buried in a mixture is to directly train distribution models, \( P(Y|M_i) \), to include the ‘typical’ effects of background sounds, either by training on noisy tokens (the ‘multicondition training’ paradigm used, for instance, in the Aurora task (Pearce, 1998)), or by synthetically combining model representations of clean targets with models of isolated interference sounds to predict the appearance of various possible forms of corruption (known variously as ‘HMM decomposition’ (Varga and Moore, 1990) or ‘parallel model combination’ (Gales and Young, 1993)). It is, however, difficult or impossible to construct a training corpus with any kind of generality: not only are an uncountable number of possible background sound objects, but when combining models the absolute signal level can no longer be simply normalized away: instead, any pair of sounds must be modeled at an enumerated range of relative levels. This combinatoric explosion results in models that are either overly broad (because a single model is being made to stand for a broad range of noise or levels), or prohibitively expensive to create and to use (because a very large number of individual models must be tested).

The alternative approach is retain \( P(X|M_i) \), the clean feature distribution model, as the basic representation of each source, but to further model the relationship between the clean features \( X \) and the compound observations \( Y \). In general, we can integrate the \( P(M_i|Y) \) term in equation 3 over the unknown values of \( X \):

\[ P(M_i|Y) = \int P(M_i, X|Y) dX = \int P(M_i|X, Y) P(X|Y) dX \]  

(4)

The first term in the integral reduces to \( P(M_i|X) \), since the value of total observation \( Y \) is immaterial given the target features \( X \). To express this in terms of our original distribution model, \( P(X|M_i) \), we can apply Bayes’ rule to this first term to give:
\[ P(M_i | Y) = P(M_i) \int P(X | M_i) \frac{P(X | Y)}{P(X)} dX \] (5)

(Note in this case that \( P(X) \) is not a constant, and cannot be dropped.) In this form, the relationship between target source features \( X \) and composite observations \( Y \) is defined by \( P(X | Y) / P(X) \), the change in the likelihood of a particular value of \( X \) given knowledge of \( Y \).

When the multidimensional distributions \( P(X|M) \) are represented as mixtures of diagonal-covariance Gaussians (GMMs), the likelihood of each mixture component can be calculated as the product of the likelihoods of the individual feature dimensions, e.g., for a mixture of \( Q \) Gaussians indexed by \( k \), over \( P \) dimensions indexed by \( j \), we have:

\[ P(X | M) = \sum_{k=1}^{Q} P(k | M) \prod_{j=1}^{P} P(x_j | k, M) \] (6)

where \( x_j \) is a scalar element of the feature vector. Assuming a similar decomposition of the prior \( P(X) \), we can use this to decompose equation 5 to give:

\[ P(M_i | Y) = P(M_i) \sum_{k=1}^{Q} P(k | M_i) \prod_{j=1}^{P} P(x_j | k, M_i) \frac{P(x_j | Y)}{P(x_j)} d x_j \] (7)

where each \( P(x_j | k, M) \) is a simple unidimensional Gaussian. The relationship between observed and target features has thus been decomposed to a likelihood change of the individual target feature elements due to the observations, \( P(x_j | Y) / P(x_j) \).

Even with this factorization, evaluating the full integral over every dimension of \( X \) will be tractable only under certain special conditions. In the ‘missing data’ approach to speech recognition (Cooke et al., 2001), it is assumed that some elements of the observation feature vector are likely to be dominated by the target sound, thereby enabling at least part of the clean-signal model to be used unmodified. This is a good match to the situation if our features are spectral energies: Many sounds concentrate their energy at any moment into a few frequency bands (such as the formants in speech), and these bands can ‘poke through’ the energy of background sounds to permit largely unobstructed observations of those parts of the spectrum. This is in contrast to the more commonly-used cepstral features, where a change in any single frequency band will, in general, change every cepstral coefficient.

Given a way to decide which observation elements reliably reflect the underlying target features, and which ones have been corrupted, we have several choices for evaluating the per-dimension integral in equation 7:

• For dimensions considered reliable, \( P(x_j | Y) \) is a Dirac delta at the assumed value, so the integral reduces to

\[ P(x_j | k, M_i) / P(x_j) |_{x_j = \hat{x}} \] (8)

• If, for a particular element, the masking due to energy from other sound sources meant that nothing could be inferred about the underlying \( x_j \), \( P(x_j | Y) / P(x_j) \) would be unity, and the integral over the complete pdf \( P(x_j | k, M) \) would also reduce to unity.

• Even if the target signal is masked at a particular frequency, we know the observed spectral energy at that frequency, and we can infer that the actual target energy is not more that this value. Thus a more accurate treatment of such dimensions is given by the ‘bounded marginalization’ approach.
(Cooke et al., 2001), where $P(x_j|Y)$ is zero for $x_j$ greater than some ceiling $x_{\text{max}}$. While the bounded distribution of $x_j$ may be difficult to express, the ratio $P(x_j|Y)/P(x_j)$ can be given a simpler form: zero for $x_j > x_{\text{max}}$ and a constant value $F$ for $x_j \leq x_{\text{max}}$, where $F$ is the normalization constant that preserves as a true pdf, i.e.:

$$F = \frac{1}{\int_{-\infty}^{x_{\text{max}}} P(x_j)dx_j}$$

which is simply a lookup of a value of the error function $\text{erf}$ when $P(x_j)$ is a Gaussian.

- More complex assumptions about the relationship between $Y$ and $x_j$ can be accommodated through other relationships. For instance, in ‘soft missing data’ (Barker et al., 2001), the true value of $x_j$ is taken to be close to the masking ceiling in regions adjacent to unmasked energy.

Thus, we see that using spectral feature models, a simple masking assumption, and some mechanism for distinguishing between masked and unmasked elements, the model likelihoods in equation 7 can be evaluated in most cases with high computational efficiency.

Preliminary Results: Figure 1 presents example results from the missing-data approach used for the standard Aurora noisy digits task (Pearce, 1998). In this task, fluent digit strings (e.g. “eight one seven three oh”) are artificially mixed with various real-world noise backgrounds (restaurant, car, airport etc.) at a range of signal-to-noise ratios (SNRs). There are two alternative training sets: the “clean” set consists only of the digits, to test how well systems can deal with completely unanticipated noise; the “multi-condition” training set includes training examples mixed with noise at a range of SNRs from clean to 5 dB.

The test set consists of distinct digit strings mixed with four noise types at seven SNRs (clean to -5 dB) for a total of 28 conditions (separate test sets include noise less similar to the noises in the multicondition set, and channel coloration). The task also specifies a ‘baseline recognizer’ built from the well-known HTK toolkit, using a standard (but optimized) set of features and parameters.

The missing-data system used spectral features (instead of the Mel cepstra of the baseline) and estimated a static background noise level from the first 100 ms of each sound file in the test set; time-frequency cells significantly above the noise floor were taken as reliable, those below were subjected to bounded marginalization, and cells whose energy was within a few dB of the estimated noise floor made a ‘soft’ contribution to overall likelihood, calculated as a linear mix of reliable and masked estimates (Barker et al., 2001).

Figure 1 shows that using the same clean-data models as the baseline recognizer, missing data recognition achieves a substantial reduction in the word error rate for higher signal-to-noise ratios, bringing performance close to that achieved by multicondition training – but, unlike the multicondition system, without any prior knowledge of the corrupting noise styles, making it far more robust to variation in test conditions.
Figure 1: Word error rate vs. signal-to-noise ratio for several approaches. “HTK clean training” uses conventional modeling and recognition, trained on clean data only. “HTK multicondition” uses conventional models, but where the training examples have been mixed with noise similar to that used in the test conditions, at similar levels. “MD Soft SNR” uses models trained on clean data only, but using a ‘soft’ variant of the missing-data recognition to do the classification.

In this system, the missing data mask, indicating which of the observed features (Y in the exposition above) are directly informative about target features (X) is derived from a simple fixed noise floor model. The outstanding question is how to identify which frequency channels should be considered as reliable, and which to treat as corrupt, in more complex conditions where the interference is anything but stationary. The next section describes using spatial and signal-structure cues to detect regions that appear to belong to a single source or voices, then section 3.3 presents speech fragment decoding as a way to combine these regions into coherent evidence to support the word-level transcription of a target voice.

**Masks from Spatial Information and Signal Characteristics:** There are two kinds of information that can be used to help identify target source features within an acoustic mixture:

- Between-channel cues, i.e. properties derived from the differences between different mics in a multi-mic signal. In general, the comparison is done in a way that cancels out the influence of the source signal itself, so, to first order, the inter-channel cues depend only on the spatial arrangement of the sound source relative to the microphones.

- Within-channels cues that can be identified in the signal from even a single microphone, and which specifically describe the source signal, rather than trying to factor it out.

Thus, between-channel cues are largely independent of the sound source’s properties (as long as it is spatially compact) and can be applied uniformly regardless of the signal. Within-channel cues on the other hand are very specific to the source: what we can do for voice would be very different that what we would do with, say, gunshot sounds.

One simple way to use between-mic cues to build a missing/present data mask is based on nonlinear
“direction-pass filtering” (Okuno et al., 2001). For each time-frequency (TF) cell in a parallel short-time Fourier analysis of all channels, a vector of between-mic time (or phase) differences and between-mic level ratios (or differences in dB) are calculated; by comparing these values with previously-learned templates from sources with known directions, a distribution of the posterior probabilities that a certain TF cell originated from any particular direction can be constructed, and thresholded to give missing-data mask. Summing across all cells gives an estimate of the spatial distribution of current sources, which can be used to choose a target direction, or the required target direction can be predetermined as part of the interface definition e.g. as directly in front of a display.

In a scenario with a close-talking mic in addition to a small array, directional information from the array is still useful in determining which energy in the close-mic channel properly originates from the target voice, and which is due to high-energy interference bleeding into that channel. Target voice energy will typically have a constant ratio between close-channel and array, and will be seen by the array to have a consistent spatial origin, depending on the location of the target speaker. High-intensity interference might have an amplitude comparable to the target voice at the close-mic, and the single channel will have no simple way to tell it apart from target; however, from the array’s point of view, the strong interference will most likely have a clearly distinct spatial origin, and a very different energy ratio to the close-channel compared to the rest of the target.

The within-channel cue most commonly used in speech separation is the harmonicity of the voiced portions of speech: vowels and similar speech sounds have an approximately periodic structure, which appears as a set of equally-spaced peaks – harmonics – in a short-time Fourier transform of sufficient resolution. Since the speech energy is concentrated in these peaks, they have a locally-maximized signal-to-noise ratio, and it has been suggested that the ‘purpose’ of these periodic episodes is to help the auditory system separate speech. Many so-called computational auditory scene analysis (CASA) systems, including (Weintraub, 1985; Cooke, 1991; Hu and Wang, 2003), are based on this approach, as are several systems not explicitly motivated by human listeners (Nakatani and Okuno, 1999). While harmonicity may be ambiguous when multiple periodic signals are present, and although a significant portion of the full speech signal is not well modeled as a harmonic spectrum, identifying clearly-displayed regions of harmonic structure can help to assemble useful target signal energy.

While harmonicity presents one of the most convenient cues to exploit, psychoacoustic experiments also suggest several other signal-intrinsic cues to help separate target from interference (for a review, see (Cooke and Ellis, 2001)). Common onset, in which energy that appears simultaneously in different frequency bands is likely to be perceived as originating from a common source, is a particularly strong principle that can explain the fusion of harmonic and nonharmonic components of many speech sounds. The very general concept of continuity, by which time-frequency regions without abrupt changes in characteristics are assumed to share a common explanation, is also important for simplifying the analysis of complex sound mixtures.

**Speech Fragment Decoding:** Speech fragment decoding (also known as multi-source decoding) is a process of matching or rejecting individual patches of time-frequency based on how well the overall collection of available data “glimpses” are consistent with previously-learned models of the complete speech signal (Barker et al., 2004).

The basic idea is to introduce the assignment of time-frequency cells to present or missing data masks as another unknown variable, and to try to solve the entire problem – both data mask, and speech recognition – in a single inference. Specifically, if we denote a particular target/interference labeling of all the time-frequency cells as a segregation S, the recognition problem (eqn 1 above) now becomes:
\[ M^* = \arg\max_{M_i} \max_S P(M_i, S|Y) \]

\[ P(M_i, S|Y) = P(M_i) \left( \int P(X|M_i, S) \frac{P(X|S, Y)}{P(X)} \, dX \right) P(S|Y) \]

Apart from this, the big difference is now that the speech recognition processes (commonly called ‘decoding’ because of its use of the Viterbi algorithm from the convolutional coding literature) in addition to (notionally) searching over all possible word strings to find the most likely match to the observations, also has to search over (an approximation to) all possible segregation hypotheses. While the basic calculations have now been defined, the \( 2^{MN} \) possible divisions of a signal consisting of M frequency bins over N timesteps into “present” and “missing” classes rapidly becomes intractable for any practical situation.

To overcome this, we make two simplifying assumptions: (1) time slices of the overall segregation mask are largely independent, so can divide the mask up into temporally separate regions, and simply combine the best solutions within each slice to get the best overall solutions, and (2) the outcome of our signal property analysis, while unable to give us a single, definitive mask (because of noise in the extracted parameters, and uncertainty over the appropriate target) does at least give us a collection of contiguous time-frequency regions – speech “fragments” – which may be confidently assumed to all originate with the same source, so a labeling of any part of such a region must extend to the whole region. Now the search space is reduced to \( N' \cdot 2^M' \), where \( N' \) is the reduced number of time slices obtained by eliminating the distinction between successive time steps that overlap the same set of (temporally extended) fragments, and \( M' \) is the number of simultaneous active (temporally overlapped) fragments – generally orders of magnitude smaller than the total number of frequency channels.

The search process then, in effect works forward, considering each speech fragment offered by the spatial and signal characteristic analysis in turn, and seeing if the overall best model likelihood (eqn 10 above) is improved or diminished by its inclusion. Hence, spurious fragments that are not consistent with the current wordstring hypothesis are automatically rejected.

Preliminary experiments, based on the simple static background noise model, and using a variety of heuristics to divide large, contiguous mask region into a more flexible collection of fragments, confirm its ability to improve performance, especially in nonstationary noise containing transients that would otherwise be misinterpreted as part of the speech (Barker et al., 2004). For example, recognizing digit
strings mixed at 5 dB SNR with factory noise (which contains many transient bursts of interference), the fragment decoding approach improves the word error rate over conventional missing data recognition from 30% to 22%, a relative improvement of better than 25%. We believe that the use of richer segmentation information, better models of segregation likelihood, and more extremely dynamic interference conditions, will make this margin of improvement much larger.

E.3. Summary of Technical Approach

To recap the preceding section, missing data recognition provides a technique for performing the match between observed signals and stored models, the essential step in speech recognition, for observations that only partially reflect the target speech due to interferences from unrelated interference such as environmental noise and incident speech. Preliminary results show that this technique rivals the best conventional noise-robust approaches even when only a crude, stationary model of background noise is used. For more dynamic noise conditions, masks indicating which regions of time-frequency are related to the target source can be derived from multichannel cues, using the differences between different mic signals to identify interference, and signal-structure cues such as harmonic and onset-time links between different frequency regions. Finally, these potential target regions are organized into consistent sets, and recognized into words, via speech fragment decoding which classifies candidate fragments as relevant to the target voice on the basis of how consistent they are with the simultaneously-developed most likely word-string hypothesis.

While we have focused on this approach as the most powerful and innovative part of this project, these techniques can also be combined with conventional approaches to feature robustness and compensation, and indeed such a combination is required if the models being tested in the speech fragment decoder are to remain comparable with the distorted speech. Thus, we will investigate combining the missing data approach with multichannel signal enhancement, robust features, and model adaptation approaches. Some of these combinations are straightforward, while others are more involved. For instance, the idea of feature mean and variance normalization cannot be applied directly to feature sets that are only partially observed; instead some other principle, such as inferring the full distribution, or normalizing to a a high percentile (or the maximum, rather than the mean) is required.

E.4. Commercialization Efforts (to be funded by WSI internal funds) and Matching Funds

Throughout the Phase I effort, we will work in concert with Navy sponsors and our industry partners, to ensure that the speech recognition technology can be seamless integrated with Naval applications and legacy military systems. To facilitate technology transfer to the Navy, we will work in Phase I to address top-level hardware and software integration issues from a systems engineering perspective. Issues such as hardware and control electronics, software architectures, hardware interfaces, manufacturability, ruggedness, and reliability will be considered in Phase I and implemented in Phase II.

Phase II Fast Track: As evidenced by myriad military and commercial opportunities in communications, controls, and signal processing, the proposed speech recognition technology holds great promise for commercialization. Accordingly, WSI will apply for a Phase II STTR Fast Track award upon receipt of the Phase I award. Because WSI has never secured a Phase II STTR contract, we are only required to obtain $188K of Fast Track cash contribution to the Phase II STTR program funding of $750K. We will work diligently in Phase I to secure this Fast Track cash match from industry and government.
E.5. Work Schedule and Task Breakdown

There is much ongoing research on enabling technologies for noise robust speech recognition. The most promising approaches will be identified and described, highlighting their potential for dependable and robust recognition in chaotic environments. The key tasks are presented below. In general they encompass characterizing the chaotic acoustic environment and reviewing current research on noise robust recognition including the components of our Speech Fragment Decoding approach.

The proposed schedule, as shown in the figure to the right, is based on seven months project duration for Phase I and another three months for the Phase I Option. Interim and final reports will be prepared and delivered to Navy sponsors. The final product of the Phase I effort will be a set of aural models and test data, an overview and descriptions of noise robust recognition technologies and a new high performance algorithm for speech recognition in chaos.

Task 1. Characterize Chaotic Aural Environments

The chaotic aural environment will be researched. Emphasis will be placed on collecting speech samples representative of chaotic environments consisting of incidental speech, room reverberation, alarms and other forms of background noise. These will be used for system testing and for developing realistic aural models capable of accurate simulation of chaotic noise scenarios as discussed in section E1. This task is expected to last three months. Expected Results: We will provide a new corpus of chaotic speech samples and new mathematical model simulations of chaotic environments.

Task 2. Speech Signal Enhancement

We will review current research on waveform and speech feature signal enhancement technologies, both single and multi-channel approaches. Acoustic beamforming and techniques for signal source separation will be reviewed among others. Performance results will be obtained from experiments performed using small vocabulary recognizers modified to implement the different speech signal enhancement schemes and the aural models developed in Task 1. This task is expected to last two months. Expected Results: Word error rate performance results for the different enhancement schemes will be reported.

Task 3. Robust Feature Representations

This task will review techniques for noise robust feature vectors such as normalized Mel-Frequency Cepstral Coefficients and Perceptual Linear Prediction. Performance results will be obtained from experiments performed using small vocabulary recognizers modified to implement the selected robust features and the aural models developed in Task 1. This task is expected to last two months. Expected Results: Word error rate performance results for the different robust features will be reported.

Task 4. Model Compensation

This task will review model compensation schemes such as maximum-likelihood linear regression (MLLR). This task is expected to last two months. Performance results will be obtained from experiments using small vocabulary recognizers modified to implement the different model
compensation schemes and the aural models developed in Task 1. **Expected Results:** Word error rate performance results for the different model compensation schemes will be reported.

**Task 5. Speech Fragment Decoding**

Based upon our prior research and expertise we are in position to propose an algorithm already drawn from current research. This task will describe the combination of missing data recognition and speech fragment decoding technologies from which our algorithm is derived. We will build in software an implementation of a small vocabulary Speech Fragment Decoding based recognizer and use it along with the aural models to measure its performance. This task is expected to last 2 months. **Expected Results:** Word error rate performance results for our Speech Fragment Decoding scheme will be reported.

**Task 6. Refinement of Our Approach (Phase I Option)**

During Phase I we will describe in detail our approach, showing the viability of the individual components and the algorithm as a whole. During the Phase I Option we will refine and develop our algorithm in detail for immediate implementation in Phase II. We will provide initial evaluation results for our approach as stated in Section E. This task will last three months. **Expected Results:** We will provide a complete description of our approach including computer models and extensive test results.

**F. RELATED WORK**

**Noise-robust Speech Recognition:** As part of a project sponsored by the European Union, the proposed Principal Scientist, Prof. Ellis invented the “Tandem” approach to acoustic modeling, which successfully combined a neural network discriminative phoneme model with a more conventional Gaussian Mixture Model-Hidden Markov Model system to achieve speech recognition in noise significantly better than either technology alone. In the 2001 Eurospeech evaluation on the standard “Aurora” noisy digits corpus, this approach achieved 50% fewer errors than the baseline, and was the best out of 20 systems presented, including efforts by major industrial research labs including AT&T, Philips, and Microsoft (Ellis & Reyes, 2001)

**Speech Separation:** Under an NSF-funded project entitled “The Machine Listener”, the proposed Principal Scientist, Prof. Ellis has been developing new statistical models to help in separating speech from interference including other voices and background noise. In one approach, an array of microphones is combined via filters whose parameters are chosen to maximize the match between the filtered output and the models present within a speech recognizer (Reyes, Raj & Ellis, 2003). In more recent work, we have considered the problem when only a single channel is available, and made the problem tractable by dividing the signal into frequency subbands (Reyes, Ellis & Jojic 2004).

**Novel Approaches to Speech Representation:** As part of the ongoing DARPA EARS program to radically reduce speech recognition errors, the proposed Principal Scientist, Prof. Ellis has been involved in several innovative front-end algorithms as part of the “Novel Approaches” spoke. By considering the information that is not preserved by the standard speech features based on 25 ms frames, but which might still be relevant, we devised a new representation for the fine temporal structure of a signal, based on the same linear-prediction equations more commonly used to model spectral structure, which offer the advantage of parametrically trading smoothness for detail (Athineos & Ellis, 2003). At the next stage of processing, we have devised a technique for selecting features across a very wide time-frequency window that are maximally informative, in an information-theoretic sense, about the speech sound class, and used this to significantly improve classification of vowels (Scanlon, Reilly & Ellis, 2003).
Mobile Phone Speech Recognition: While employed at Tellabs the proposed Principal Investigator, Mr. Morris Small, worked on a hands-free mobile phone kit that was developed for a Japanese auto parts manufacturer that enabled voice controlled mobile phone operation. Using this car kit a driver could completely control their mobile phone. Mr. Small worked with the developers of the recognition engine to specify interfaces between the kit and the speech recognition engine. He implemented voice commands such as dial, speed dial, redial, name dial and on-hook. He integrated the recognizer with other Tellabs voice quality enhancement software and conducted tests with the customer.

Noise Reduction: The proposed Principal Investigator, Mr. Small, has worked on conventional noise reduction schemes for telecommunications applications. Another of Tellabs’ products that Mr. Small contributed to was a floating point implementation of a spectral subtraction type noise canceller. Mr. Small debugged and tested its hardware implementation. Mr. Small also developed software for a simple fixed point implementation of a background noise minimizer based on signal strength and integrated it as an additional feature offered with a telecommunications echo canceller product.

F.1 Previous WSI Customers and Projects

Audio Conference Server: Washington Software assisted Tellabs in building the Consortium Conference Server, which was a real-time distributed conference system that allowed up to 96 people in each conference. The system used Microsoft Windows 2000 server. The hardware platform was ISA passive backplane chassis, using Multi-Vendor Integration protocol (MVIP-90) interface. The system provided clear audio call quality automatically by using dedicated echo canceling boards, dynamic port adaptation, automatic gain control, voice level equalization, full duplex audio, and removing intrusive background noise. The system also provided an Interactive Voice Response (IVR) interface for users to dial-in, join a conference, and schedule a conference. Users could administer the conference server through the web or distributed software clients.

Information Technology: Washington Software is an IT solution provider offering IT consulting and applied Research & Development (R&D) consulting services. Our IT consulting service includes application development in financial systems, data analysis applications, computer telephony integration, and e-business systems. The applied R&D consulting service includes the evaluation and prototyping of emerging technologies, and recommending a revolutionary solution to our clients. Our iterative and adaptive development methodology successfully reduces technical risks associated with development projects. Our effective methodology and veteran engineers distinguish us from our competition. Our mission is to realize our clients’ vision by providing best value, state-of-the-art solutions that satisfy their business needs. Our clients include Lockheed Martin, Abt Associates, Guardian Insurance, Tellabs, IBM (Subcontract), Caterpillar Financial (Subcontract), Department of Labor (Subcontract), and Fannie Mae (Subcontract).

G. RELATIONSHIP WITH FUTURE RESEARCH OR RESEARCH AND DEVELOPMENT

The proposed research team, especially the proposed Principal Investigator and Principal Scientist, Mr. Morris Small and Dr. Dan Ellis, respectively, have enjoyed long-standing collaborative research and development relationships with industry and government research laboratories. For this specific project, WSI has established a strategic alliance with Tellabs that possess the entire range of capabilities, resources, and production assets to assist WSI to prototype, integrate, test, and produce the speech recognition technology for Navy and commercial applications. Throughout Phase I and in preparation for Phase II, we will work closely with Navy sponsors and our industry contacts to carefully address full-scale systems integration issues and engineering challenges from a end-user perspective. Central to our success in potential transition to the Navy fleet is the ability to configure the speech recognition
technology as retro-fit system and to ruggedized and package the technology for the field use environment where performance over a wide range of conditions and standard maintenance practice requirements must be met.

In Phase II, appropriate modifications to the Phase I prototype system will be made for full-scale implementation. Systematic testing will be conducted to simulate tactical operating conditions such that the speech recognition can be ruggedized for field use. The processing algorithm will be downloaded onto a digital signal processor (DSP) board and combined with the on-line controller circuitry, such that the entire system can be combined into a modular unit. In addition to full-scale device development, full-scale loading tests will be conducted in the laboratory and in a simulated combat environment agreeable to the sponsor. The end product of the Phase II technology transition effort will be a proven speech recognition technology and associated control hardware/software system that can be potentially retro-fitted and integrated with fielded Navy systems to provide improved communications operations under tactical conditions.

H. COMMERCIALIZATION STRATEGY

For the defense market, WSI’s speech recognition technology will be an integrated software/hardware product that can be licensed for manufacture to our strategic manufacturing partner, Tellabs, or a military systems integrator such as Lockheed Martin or IBM, depending on the market being addressed. As described in Section F, WSI has been serving the commercial market over the past 5 years. Because WSI and Columbia University has already established a presence in the acoustic research and products markets through our existing customers, we plan to leverage these marketing outlets and offer speech recognition systems for enhanced performance where conventional technology have not been successful from performance and cost perspective. In Phase II, WSI will finalize a formal partnership with a strategic manufacturing partner, who will produce the speech recognition systems specific to a commercial application. WSI will perform preliminary subsystems integration of the speech recognition systems, develop software upgrades, and conduct direct marketing and sales of the product to the end-use through our internal resources.

COMPETITION AND TECHNOLOGY RISK: We are unaware of any commercial competitive technology that currently exists that captures the features, advantages, and benefits of the proposed technology. Navy STTR funding and the combined capabilities of our proposed research team will afford us the opportunity to prototype, develop, and field the speech recognition prior to our potential competitors who have focused on traditional noise cancellation systems that are much more expensive and complex. We view the proposed technology development effort as low risk to the sponsors because our system leverages existing capabilities, resources, and expertise of the entire research team. Due to its simplicity, the proposed technology is easily producible in large numbers with reliability at low cost.

TRACK RECORD AND INVESTMENT COMMITMENT: Historically we have an excellent track record leveraging government investment to produce commercial software products and services. We have successfully leveraged government funds for revenues toward software development. WSI will fund all marketing efforts directly from our internal funds. In Phase I, WSI plans to commit $15k and $20k for commercialization and to augment STTR resources for prototype testing, respectively. Based on the past and planned financial support for the development of this technology by WSI, we believe that the investment of these resources demonstrates the commitment of our entire team to this effort.

<table>
<thead>
<tr>
<th></th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-SBIR Revenues</td>
<td>$350k</td>
<td>$1,000k</td>
<td>$3,500k</td>
</tr>
<tr>
<td>Employees</td>
<td>0</td>
<td>+1</td>
<td>+3</td>
</tr>
</tbody>
</table>

INTELLECTUAL PROPERTY: As a matter of standard practice, WSI protects our IP, trademarks, and trade secrets through rigid enforcement of non-disclosure agreements with
potential strategic partners, prompt and accurate technology disclosures, and where appropriate, patent disclosures and copyright applications. We are experienced in business practices, product development, manufacturing, and domestic/international sales. WSI works with local IP counsel on a retainer basis to ensure appropriate protection of our technologies and practices. WSI is certified as a 8(a) disadvantaged small business.

**BENEFITS SCHEDULE:** WSI expects a selling price of approximately $20,000-$35,000 for the speech recognition system, depending on volume. The revenues expected above are based upon the assumption that we will outfit 10 navy systems in year 1, 30 systems in year 2, and 150 systems in year 3.

I. KEY PERSONNEL

**Mr. Morris Small, Senior Scientist at Washington Software, Inc.,** will serve as **Principal Investigator**. Mr. Small earned his M.S. degree in Electrical Engineering from Northeastern University, Boston, MA., and his B.S. degree in Electrical Engineering from George Washington University. He has over twenty years experience researching and developing signal processing applications and systems. He began his career with the development of a speech bandwidth compression system based on an aural model optimized filter. Throughout his career he has engaged in numerous research and development activities for voice quality enhancement products, sometimes developing computer models using MATLAB or C, and at other times developing and integrating product software. Among his research interests are techniques for background noise reduction. He has worked on both fixed point and floating point noise reduction products, developing algorithms as necessary and integrating them into telecommunications products. At Tellabs, his previous place of employment, he integrated a continuous speech recognition engine capable of both independent and dependent speech recognition into a mobile phone.

**Dr. Daniel P.W. Ellis, Assistant Professor in the Department of Electrical Engineering at Columbia University,** will serve as **Principal Scientist for speech recognition algorithm development**. Dr. Ellis earned his bachelor's in Electrical and Information Sciences from Cambridge University, England in 1986, and completed his graduate education at MIT, earning an SM in 1992 and Ph.D. in 1996. He then spent four years as a research scientist working on speech recognition at the International Computer Science Institute in Berkeley CA before joining the faculty at Columbia in 2000, where he leads the Laboratory for Recognition and Organization of Speech and Audio. Dr. Ellis's research interests encompass applications of signal processing, pattern recognition, and machine learning to extract all kinds of information from acoustic signals including speech, music, and environmental noise, with the separation of overlapping sound sources a particular focus. His achievements include development of the best-performing speech recognition system presented at the 2001 "Eurospeech Special Event on Noise Robustness", and the development of the SPRACHcore speech recognition software package in use at many research labs around the world. Current funding for Dr. Ellis comes from the NSF (including a prestigious CAREER fellowship) and from DARPA under its EARS Novel Approaches program to develop radical solutions to current speech recognition challenges. Dr. Ellis is the author of more than 50 peer-reviewed papers, and has been an organizer of several recent and upcoming professional meetings including the NSF Workshop on Speech Separation (November 2003) and the ISCA Research Workshop on Statistical and Perceptual Audio Processing (October 2004). Since 1993, cognitive processing in sound, which currently has more than 1200 participants.

**Selected Refereed Journal and Conference Publications**


**Dr. Peter C. Chen, Director of Technology at Washington Software, Inc.,** will serve as Principal Specialist for Systems Integration and Testing. Dr. Chen received his B.S., M.S., and Ph.D. degrees in Aerospace Engineering from the University of Maryland at College Park in 1990, 1993, and 1996, respectively. Over the past 14 years, Dr. Chen’s responsibilities have included technology development, establishing strategic alliances, marketing, business development, and product commercialization with strategic partners. He has directed and served as Principal Investigator and Principal Specialist for myriad Department of Defense, commercial, and industrial technology programs. Over the past six years, Dr. Chen has served as Principal Investigator and Principal Specialist for 15 Phase I SBIR/STTR programs, 5 Phase II SBIR programs, and 2 Phase III SBIR projects. Dr. Chen holds five United States and International patents on advanced technologies developed for active vibration control, optical sensor systems, weapons stabilization, and health monitoring applications. Past sponsors of Dr. Chen in the Department of Defense include The Defense Advanced Research Projects Agency (DARPA), Naval Undersea Warfare Center, Division Newport, Naval Surface Warfare Center, Carderock and Port Hueneme Divisions, Office of Naval Research, Naval Air Systems Command, Naval Sea Systems Command, Army Research Office, Army Aberdeen Proving Ground, Wright-Patterson Air Force Base, Air Force Arnold Engineering Research Center, Army Redstone Arsenal, Army Research Office, NASA (Ames, Langley, and Glenn Research Centers), and the Office of the Secretary of Defense. Commercial customers include Honeywell Engines and Systems, Northrop Grumman Corporation, The Boeing Company, the McDonnell Douglas Helicopter Company, and Lord Corporation.

Dr. Chen’s areas of expertise include aerospace systems, optical systems, adaptive materials, optical fiber sensors, unmanned aerial vehicles, structural health monitoring, and systems integration. He has authored and presented 30 technical papers in professional conferences and 5 archival journals related to his research. Dr. Chen has served on the Program and Technical Committees of the International Society for...
Optical Engineering’s (SPIE) (Smart Structures and Materials and Industrial and Commercial Applications of Smart Structures) and the American Society of Mechanical Engineers since 1996 and 1997, respectively. He is also a technical reviewer for the Journal of Intelligent Materials Systems and Structures, and the Army Research Office. Dr. Chen served as a contributing author in Smart Materials Technologies for the American Institute of Aeronautics and Astronautics’ publication, Year in Aerospace 1998 and The Encyclopedia of Smart Structures. Dr. Chen’s development of a smart rotor system was awarded the “Outstanding Contribution to the Engineering Profession,” by the McDonnell Douglas Company and the American Society of Mechanical Engineers in 1996. He has authored a chapter in the American Institute of Astronautics and Aeronautics’ (AIAA) Progress in Astronautics and Astronautics Series of publications, entitled “Advances in Adaptive Structures and Materials,” based on his work in adaptive rotor controls. Dr. Chen enjoys a long-standing professional relationship with UMD. Dr. Chen is cleared to the Secret level by the Defense Industrial Security Clearance Office (DISCO).

Selected Refereed Journal and Conference Publications


J. FACILITIES AND EQUIPMENT

Washington Software, Inc. (WSI) is a financially stable, small business that provides technical computer consultation, with SBA 8(a) and SDB certification, in the Metropolitan Washington D.C. area. Since 1998, we have provided IT Consulting and Applied Research & Development services to government, financial, and telecom companies. Our engineers specialize in speech signal processing research,
algorithm development and coding; transforming a product/project idea to a complete solution; integrating state-of-the-art technologies; product development life-cycle methodologies. The Washington Software headquarters facility is located in Gaithersburg, MD. This facility includes office equipment, laptops, computer workstations, servers, network of high performance personal computers, and microphones required for the proposed program. The facility has computers running different operating systems, such as different versions of Windows, Linux, and Solaris. We have state of the art design software that models and designs the system interfaces, architecture, and software/hardware component interface of the system. We have compilers for different programming languages, such as Java, C, C++, Visual Basic, and Smalltalk. We also have load testing software that test our system by simulating the work load of many end-users in real life production environment. The facility has high speed Internet connection protected by a corporate firewall.

The Laboratory for Recognition and Organization of Speech and Audio (LabROSA) at The Columbia University, (http://labrosa.ee.columbia.edu/) was founded by the proposed Principal Scientist, Professor Dan Ellis, when he joined the Electrical Engineering Department of Columbia University in 2000. In addition to office space, the lab occupies a 750 sq.ft. space in the new Schapiro Center for Engineering and Physical Sciences on the north edge of the main Columbia Campus. The lab is fully equipped with a farm of Linux workstations (up to 3 GHz) and a Terabyte filesystem, and is directly connected to Columbia’s main internet backbone. Specialized equipment includes a range of professional and semi-pro audio recording and playback equipment, including several high-quality portable direct-to-disk digital recorders with matching microphones that have been used for field recordings used to investigate “real-world” acoustic environments.

LabROSA currently hosts eight graduate student researchers funded from various sources including industrial grants, National Science Foundation projects (including one on analyzing speech in natural recordings of real meetings, and one on using missing data techniques to analyze real-world soundtracks), and DARPA via participation in the Novel Approaches track of the current Effective, Affordable Real-world Speech-to-text (EARS) project, the main DARPA project aimed at improving speech recognition.

K. SUBCONTRACTORS AND CONSULTANTS

Statement of Participation: LabROSA at Columbia University will be available at the times required for the purposes and extent of the effort described in this proposal:

Dr. Dan Ellis, Founder and Director LabROSA

L. PRIOR, CURRENT, OR PENDING SUPPORT OF SIMILAR PROPOSALS OR AWARDS

WSI has no prior, current, or pending proposals or support for the proposed development.
REFERENCES


