Results of the Phase I SBIR/NSF Project.

[*Note*: A more detailed version of Phase I results is given in our Phase I Final Report].

The general problem category addressed by our Phase I SBIR project concerned the so-called *cocktail party problem*, which has long been recognized as an important but difficult speech processing/understanding problem. It arises when we wish to process single-microphone (monaural) or multi-microphone (multi-aural) data collected in a reverberant environment in order to separate a speaker's voice from other voices and sounds in the background. Obviously, the human brain is able to perform such processing even when the background voices and sounds are quite loud, but the important question is whether a machine can be endowed with a similar capability.

By innovatively combining state-of-the-art artificial intelligence, signal processing for voice DNA analysis, machine learning, and broadcast studio techniques, Yobe Inc. has developed a novel cocktail party technology we refer to as ***Yobeization***. While this technology has not solved the cocktail party problem in its entire generality, it has made significant inroads beyond currently available solutions for practical applications of interest such as audio surveillance, hearing aids, voice authentication and identification, and voice communication for the Internet-of-Things (IoT). In its most sophisticated form, our Yobeization technology is designed for working on two or more microphone inputs to separate (on the basis of voice DNA) two-person conversations from the accompanying noise in everyday environments such as cafes, offices, homes, and street corners. Before the start of the Phase I SBIR project, we had already implemented and tested a MATLAB beta version of the *Yobe Engine* for Yobeization of desired voices in noisy environments. Given this context, our Phase I SBIR proposal set out the following research and development goal:

"Our main objective in Phase I of this SBIR is to systematically ***optimize the conversation-enhancement performance of the current version of Yobeization with respect to its use as a frontend process for standard state-of-the-art speaker and speech recognition systems***. This objective is necessitated by our desire to broaden the set of applications (beyond hearing aids and audio surveillance) for which the conversation-enhancement capability of *Yobeization* has been optimized before embarking upon a commercial hardening and productization process.

1.1 Speech Recognition Terms and Standards: Once the *Yobe Engine* has extracted the desired voice from the noisy environment, the extracted voice is passed on to a deep neural network (DNN) based ASR (*Cubic*) provided by Cobalt, Inc. The performance is calculated using the industry standard Word Error Rate (WER) method in which the word error rate is based upon the number of incorrect words and guesses divided by the number of words in a reference file. A reference file is the script used in the recording and the hypothesis file is the output of the ASR. These two files are compared and 4 situations can arise for each word. A word can match in both files (Correct), it can be replaced by another word in the hypothesis file (Substitution), a new word can be added to the hypothesis file (Insertion), or a word can be missing in the hypothesis file (Deletion). The corresponding Word Accuracy Rate (WAR) rate is defined as:

$$WAR\ = 1 - \frac{\#Deletion + \#Insertions + \#Substitutions}{\#Correct + \#Deleted + \#Substitutions}$$

The # sign denotes the total number of words in a certain category. At Yobe, we have used the *Cubic* ASR by Cobalt (http://www.cobaltspeech.com/) and the program *Sclite* (http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm) is used to compare the reference scripts and to calculate the word accuracy rate.

1.2 Speaker Recognition Terms and Standards: The form of speaker recognition we used in our Phase I project to study the effects of Yobe technology is known as the *speaker authentication* task. Given a voice recording and a *claimant* for the voice, the task is to either verify that the voice belongs to the claimant or to reject the claimant as an *impostor*. The *claimant rejection rate* is the probability that the claimant will be rejected by the authentication system given that he/she actually produced the recording. The *impostor acceptance rate* is the probability that an impostor who actually did not produce the recording will be accepted by the authentication system as the claimant. The technique we used for carrying out the speaker authentication task on recordings in our databases is known as GMM-UBM (Gaussian Mixture Model – Universal Background Model). This technique has been extensively discussed in:

1. Speaker Verification Using Adapted Gaussian Mixture Models (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.117.338&rep=rep1&type=pdf)
2. Robust Text-Independent Speaker Identification using Gaussian Mixture Models (http://ieeexplore.ieee.org/document/365379/)

1.3 Noisy Database Basics: Our noisy database is a total of 38,991 seconds worth of recordings, split among approximately 2000 files, each of 15 seconds duration. The noisy database is further partitioned into 2 sub-categories: Indoor Spaces (22,703 seconds) and Outdoor Spaces (16,288 seconds). The environment classification is as follows:

	Ambient Noise Sources	Directional Noise Sources	Moving Noise Sources	Examples
Indoor Spaces	Present	Dominant	Nominal	Conference rooms, Offices
Outdoor Spaces	Dominant	Present	Present	Cafeteria, Coffee Shops, Roadside

In the above table, a source is considered "Dominant" if it is the major source of noise in a recording. It is considered "Present" if it is typically expected to be present in a recording but it is not the dominant source of noise in the recording. A source is considered "Nominal" if it is typically not present in a recording but when it is, it is not dominating. The adult, native English speakers in the recordings followed one of 257 scripts (average duration of each script is about 70 seconds) that were selected to ensure that they were not outside the vocabulary range of the ASR system from Cobalt Inc. ***The average Word Accuracy Rate (WAR) produced by the Cobalt Inc.'s ASR Cubic (without the aid of Yobeization) across the entire noisy database was 26.8%***.

1.4 Benchmarking of Yobeization with controlled noise**:** To benchmark the performance of Yobeization as a function of SNR, we added controlled levels of ambient and directional noise to the recordings from the Noiseless database. In Figure 1, we show the ASR results with and without the use of Yobeization. The ***blue*** curve represents the Word Accuracy Rate performance of Cobalt ASR ***without Yobeization***, while the ***orange*** curve represents the Word Accuracy Rate performance of Cobalt ASR *with Yobeization* (the vertical bars at various points represent standard deviations). It is clear that the Yobeization makes a significant difference in word accuracy rates for controlled noise.

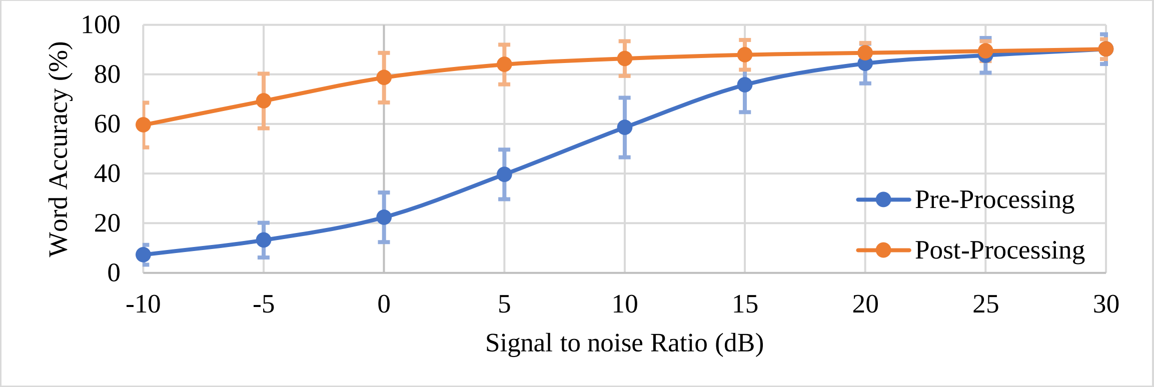

Figure 1: Word accuracy rate vs. Signal to noise ratio

1.5 Speech Recognition Evaluation of Yobeization on Noisy Database (Aim 3 of Phase I): In this section, we describe our results for applying Yobeization to improve the performance of ASR on our noisy recordings database. In Figures 2 and 3, the Word Accuracy performance is translated to an "After Yobeization SNR" and a "Before Yobeization SNR" by utilizing the benchmarking data (Section 5 of this report) for controlled mixed-noise scenarios (the vertical bars at various points represent standard deviations). Figure 2 is for the Indoor Spaces portion of the Noisy Recordings Database while Figure 3 is for the Outdoor Spaces portion of the Noisy Recordings Database. We can see from Figure 6 and Figure 7 that Yobeization leads to an improvement in SNR of approximately 20 dB on average.

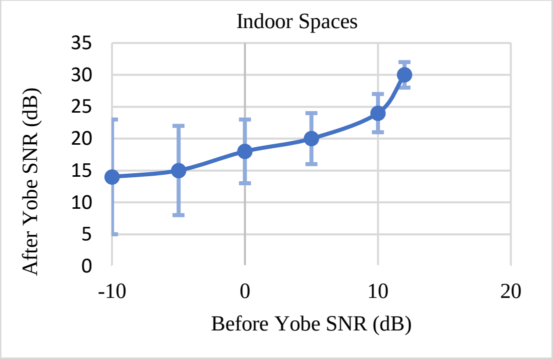

Figure 2

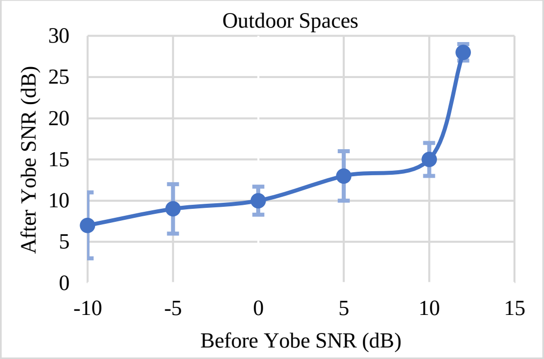

Figure 3

To further investigate the efficacy of Yobe technology for speech recognition in noisy environments, we also tested it when the IBM Watson speech-to-text program performs the speech recognition. The IBM Watson speech-to-text program is available at https://www.ibm.com/watson/services/speech-to-text/. As illustrated in Figure 4 for indoor scenarios and in Figure 5 for outdoor scenarios, Yobeization improves the accuracy of IBM Watson ASR to an even greater extent than it does for the Cobalt's Cubic ASR. This results from the fact that the Watson speech recognizer has more extensively trained acoustic and language models.

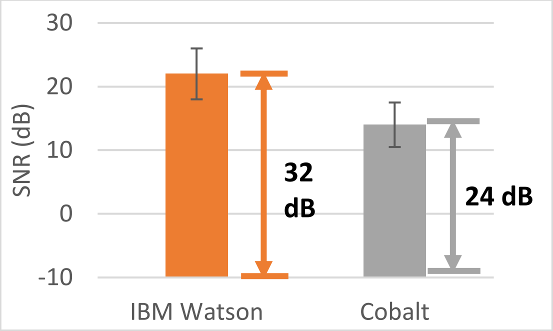

Figure 4: After Yobeization of -10dB SNR files, IBM Watson gave accuracy equivalent to raising the SNR by 32 dB while Cobalt 's *Cubic* accuracy raised the SNR by 24dB. Bottom of each bar is the starting SNR and the top represents the post-Yobe SNR.

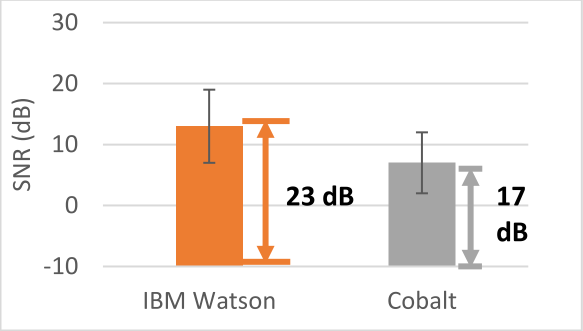

Figure 5: After Yobeization of -10dB SNR files, IBM Watson gave accuracy equivalent to raising the SNR by 23 dB while Cobalt 's *Cubic* accuracy raised the SNR by 17 dB. Bottom of each bar is the starting SNR and the top represents the post-Yobe SNR.

1.6 Speaker Authentication Evaluation of Yobeization on Noisy Database (Aim 2 of Phase I): In this section, we report the results of evaluating how much Yobeization of the Noisy Database improves the speaker authentication task within speaker recognition. In Figure 6, we show the performance of the speaker authentication task of speaker recognition in the case of Indoor Spaces. The blue curve shows the performance obtained when speaker authentication is carried out directly on the noisy recordings without utilizing Yobeization. The dashed orange curve shows the performance obtained when speaker authentication is carried out on Yobeized versions of the noisy recordings. It is clear that for any impostor acceptance rate, the claimant rejection rate is an order of magnitude lower when Yobeization is used on noisy database.

In Figure 7, we show the performance of the speaker authentication task in the case of Outdoor Spaces. The blue curve shows the performance obtained when speaker authentication is carried out directly on the noisy recordings without utilizing Yobeization. The dashed orange curve shows the performance obtained when speaker authentication is carried out on Yobeized versions of the noisy recordings. Once again, it is clear that for any impostor acceptance rate, the claimant rejection rate is an order of magnitude lower when Yobeization is used on the noisy database.

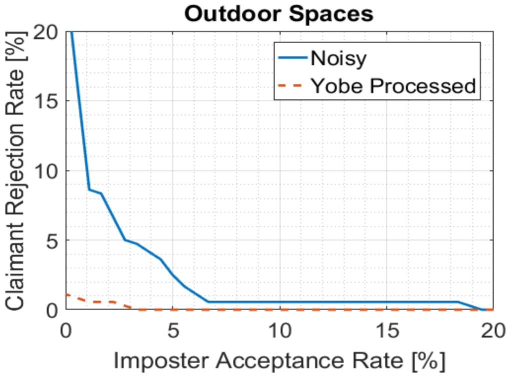

Figure 6

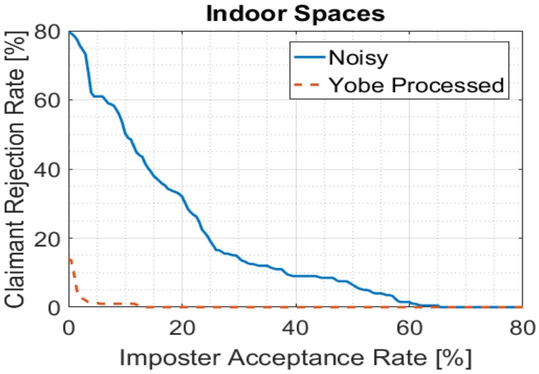

Figure 7

2. Conclusion

In the Phase I SBIR project that we have just concluded, we have definitively demonstrated that our Yobeization process (as currently encoded in MATLAB) is able to effectively raise the SNR of real-world noisy recordings by approximately 20 dB for the purposes of speech recognition and/or speaker authentication applications. This means that Yobeization brings real-world noisy recordings with SNR's in the "impossible range" of -10 dB to 0 dB SNR (beyond the reach of modern speech and speaker recognizers) into the realm of SNR's above +10 dB, where those same recognizers are considered effective for most practical purposes. As envisaged in our original Phase I proposal, the achievement of this objective opens the way for the commercial hardening and productization of the Yobeization process. The stage is now set for converting our current MATLAB code for Yobeization into a modular "Yobe Engine" in hardened C++ code that is also optimized for real-time performance on a variety of hardware platforms. While there can be a variety of different verticals for which the Yobe Engine can be specialized, our objective in Phase II will be to specialize it for a "Voice User Interface" or VU/I. Such a VU/I would provide the capacity for identifying voice biometrics from noisy microphone signals and improve the SNR with respect to the desired voice signals. The resulting voice signals and their cleaned up voice biometrics may then be fed to state-of-the-art third-party speech and speaker recognizers which are only able to operate at relatively high SNR's. Our Phase I SBIR project has firmly established the practicality of such a vision for Yobeization and its implementation in Phase II.