Active Optical Source Modulation and Filtering for Coherent Detection of Light Scatter

B. Michael Kanack
University of Kentucky
Pigman College of Engineering - Paducah Campus
4810 Alben Barkley Drive
Paducah, KY 42001

ABSTRACT
Since Stanley Kubrick's, epic 1968 science fiction film, "2001: A Space Odyssey", computer comprehension of visual human speech directly to words has been imagined. There have been hundreds of publications and book chapters devoted to this topic, including multiple review articles. Examples are Gaikwad et al. [1] and Battista et al. [2]. Most papers depend on Artificial Intelligence based upon Hidden Markov Models and other camera techniques capturing and analyzing visual motion in time of video frames. Some notable exceptions are direct monitoring of the laryngeal, Jorgensen [3] and embedding magnets, Gilbert et al. [4].

In this paper, a novel technique is used to improve the capture of these motions and process the utterance of an entire word into a Composite Ensemble Phonemes Viseme (CEPV). Described is the improvement in the optical sampling technique. This is accomplished both with an active optical source and a high speed camera. Subsequent digital signal processing allows for compression of time domain scattering of light of an utterance into a unique image or matrix of numbers, representing a word which can be compared to other images within a list to interpret meaning.

Keywords
Lip Reading, Discrete Fourier Transform, Laser Illumination, Covert Communications, Urban Security/Warfare.

1. INTRODUCTION
An application where non-verbal communication is at a premium is tactical communication within a security team. A comprehensive system for **C**overt **A**dvanced **C**ommunications for **T**actical **U**rban **S**ecurity (CACTUS) has been created for this task. CACTUS proposes a helmet mounted Advanced Optical Processing Apparatus to synthesize words directly from facial motion of the user, primarily lip and chin motion. The user has a minimal learning curve since normal facial motion to form the desired word is used. Once the CACTUS system has sampled the user's input, CACTUS will directly translate the facial motion into text. The text can thereby be transmitted to the user's team. Alternatively, the text can be translated into sound for radio transmission using a voice synthesizer. Finally, the user's own voice can be previously recorded and transmitted by using the text to cross reference an audio look-up table. Other facets of the proposed system are that no line of sight is required for the user to communicate with the team such as is required for visual hand signaling. CACTUS will integrate direct multilingual translation to facilitate multinational tactical communication within a security force. CACTUS can also be used for high noise environments such as flight line operations. CACTUS operates as a normal communication headset out of stealth mode.

2. SYSTEM PROPERTIES
Tactical teams require a high level of reliability, however, the entire number of words can be very low. The system is designed to capture these non-verbal expressions of facial movements which would be an utterance of tactical commands. These silent gestures are transmitted to the team members via previously recorded utterances or alternatively a voice synthesizer

2.1 Light Source Characteristics
Operation at night, in bright changing sunlight or rapid changing glare all represent challenges to optically sampling utterances from a subject. Moreover, publications typically sample lips at a normal angle, whereas for the average subject, symmetry can be assumed. Finally, location of the sampling area is not fixed, with respect to the camera or the illumination.

Here we mitigate all of these problems by 1) use of an active optical source, proposed to be laser; 2) using a narrow band optical illumination; 3) fixing the optical source at a known proximity; and 4) fixing the angle of the optical source to the sampling area to create a major light scattering angle, changing the usual regular symmetry.

An active optical illumination source is obviously required for night operations or where the ambient light is not sufficient. Moreover, using single frequency optical source allows for camera sampling to have a uniform, continuous optical source by using a complementary narrowband optical filter. It is this narrowband optical filter at the sampling device which mitigates optical glare and changing illumination characteristics. Finally, modulation of the optical source to enable coherent detection of the scattering of light from the surface, is proposed.

AlGaAs lasers are proposed to be the ideal eye safe laser with low power and a dispersed beam for operation within the CACTUS system. AlGaAs lasers operate from 650 to 980 nm and are readily available as a low cost option. Charge Coupled devices also operate in the range of 400 to 1050 nm. Frame rates routinely exceed 1000 fps. No color filters are required for operation, just a complementary narrowband optical filter for the laser wavelength.

2.2 Image Capture Requirements
Most lip reading publications use common video sampling down to 25 fps, Assael et al. [5] and Patterson et al. [6] employing ambient light. In addition to the optical filtering to present a more consistent optical illumination, independent of the environment, a high speed camera is used. Use of a high speed camera, at up to 1000 fps is two fold; 1) over sampling improves the signal to noise ratio of the sample site and 2)

allows for coherent detection of scattering due to the modulation of the light source.

In addition, the image capture device or camera, is also affixed relative to the sample area at a complementary angle to the active optical source. This capture angle of 90 degrees is optimal to light scatter from the region of interest.

The modulation rate of the optical source and optical sampling range will vary for capturing light scatter depending upon the length of words and the number of phonemes used to form such words. In English, the number of phonemes is 44, Bizzocchi [7] and the average number of phonemes in a word are about 6, Marian et al. [8]. These numbers vary with languages. An extensive review of linguistics is out of the scope of this paper. The intent is a general purpose, language independent sampling technique and therefore, over sampling is employed.

2.3 Data Analysis

Python programming language was used for the data analysis. The video stream of the utterance of a word was broken down into thousands of individual frames using OpenCV. Following, Python Imaging Library (PIL) was used to extract each X,Y pixel in the entire range of frames as a function of the sampling frequency. This data was further processed using Numpy's Discrete Fourier Transform (DFT). The amplitude of the fundamental frequency of the light source modulation of the X,Y coordinate was extracted. This process was completed for the entire range of pixels for the utterance of the CEPV resulting in a single value for that pixel, compressing the data.

2.4 Digital Signal Processing

The modulation of the light source, and the subsequent high speed capture light scatter, allow for time domain analysis between frames as opposed to spatial analysis of a given series of frames. Selection of the modulation frequency and the sampling frequency will depend upon the rapidity of the spatial variation, as previously mentioned. It is desirable to maintain an optical sampling rate which exceeds the light modulation rate by at least a factor of two to ensure operation in the first Nyquist Zone, preventing aliasing.

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-i2\pi kn/N} \quad (1)$$

$$|X_k| = \sqrt{[\mathrm{Re}(X_k)]^2 + [\mathrm{Im}(X_k)]^2} \quad (2)$$

where:

X_k: The k-th frequency bin output

x_n: The n-th input sample in the time domain

N: Total number of samples

k: Frequency index

It is the approach to track each pixel in time, not process instant frame spatial characteristics. The scatter of each pixel through the utterance of a word will have a particular ensemble modulation as the phonemes are expressed. The variation of the amplitude of each individual pixel, within the frame, for the duration of the utterance is found. The ensemble amplitude may simply be calculated using the Discrete Fourier Transform (DFT) of the pixel scatter as a function of time (eq. 1 and 2, above). It is each pixel ensemble, within the whole frame that makes the composite ensemble phoneme viseme or CEPV.

Although the DFT has been used, if the modulation frequency is fixed for the desired sampling phenomena, a more simple, pixel by pixel narrowband filter with a square law detector could be used. Also, an actual digital filter, such as a Finite Impulse Response (FIR) or an Infinite Impulse Response (IIR) filter would be effective to capture the modulation of the electrical frequency of the light. Any of these implementations could be realized in a Field Programmable Gate Array (FPGA) to process the large amount of data being created by over sampling.

3. EXAMPLE RESULTS

The experimental set up consisted of an active optical emitter and high speed camera. Since the subject of observation included a human, a darken room with common white light illumination was used. Data was taken at several optical modulation rates and frame rates, but the goal of over sampling to improve signal to noise ratio resulted in an optical modulation rate of 75 Hz and a video sampling rate of 1000 fps. The modulation of the optical source was at a 50% duty cycle.

3.1 Time Domain Scatter Results

Figure 1 is an example of a single pixel tracked through an entire utterance of a word. The inset (a) displays the changes in scattering of the light source as it is modulated between the on and off state with changes in scatter. Likewise, inset (b) displays the nearly uniform off state of the modulation.

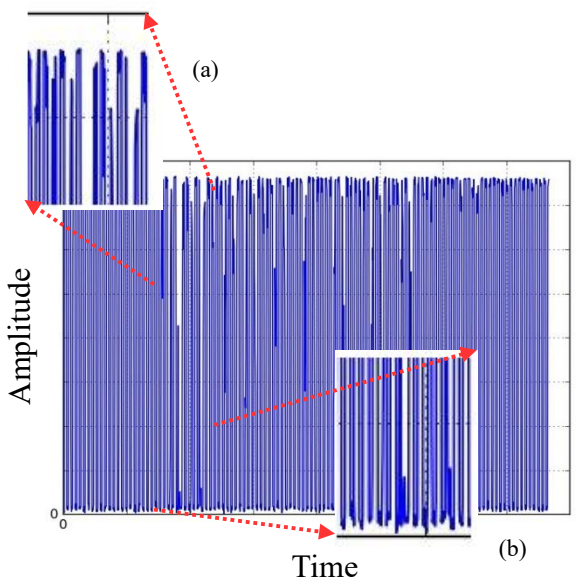


Figure 1. Time Domain Single Pixel Variation of Scatter

Figure 2 is the amplitude of the DFT of the time domain signal. The main peak (a) is representative of the ensemble phoneme pixel for a location within an entire utterance. The third and fifth odd harmonic frequencies (b) and (c) are also well represented in the DFT. Minor peaks are intermodulation characteristics resulting from detector nonlinearities, Kanack [8].

3.2 Composite Phoneme Viseme

Figure 3 is a false color image of a Composite Ensemble Phoneme Viseme. The color red represents a high level of scatter attenuation due to the movement of the lips during the

word utterance, as the individual pixel scatter changes with motion. As green moves into blue, the change in scatter has low attenuation. The amount of raw data is approximately 5 MB and has been reduced to 10 kB, a compression of 500 times.

Figure 2. Frequency Domain Single Pixel Amplitude

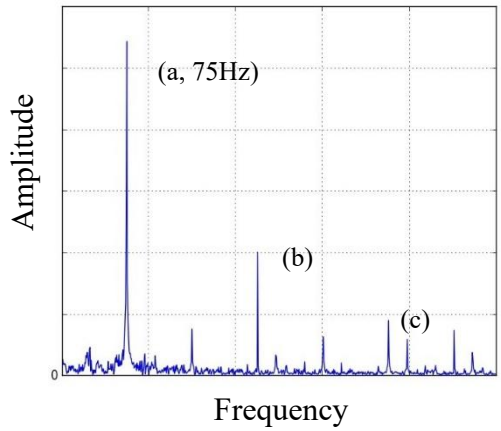
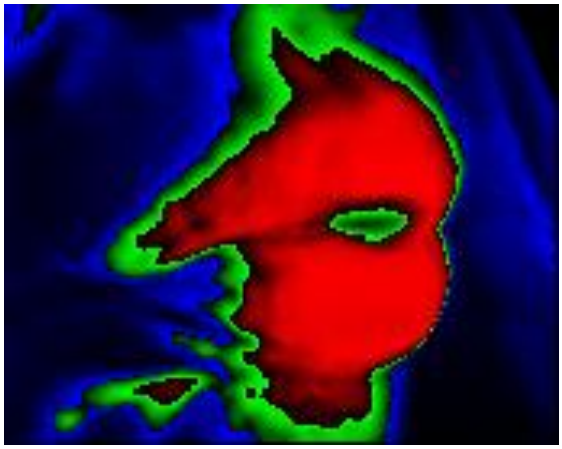


Figure 3. False Color Image of CEPV



3.3 Confusion Matrix Analysis

A CEPV "Dictionary" of words was created that would be useful to tactical teams. Words such as "UP", "DOWN", "RIGHT", "LEFT" and "FIRE" were sampled four times using the optical sampling system. An example set of eight words were used. This tactical "Dictionary" could change by the iterative field programmable nature of the headset depending upon the presumed tactical situation.

For each word, the data was conditioned into the composite ensemble phoneme viseme. Each utterance was normalized after summing the Red, Green and Blue pixel values of the color high speed camera. This was completed since white light illumination in a dimmed room was used as opposed to the proposed eye-safe laser source. Following normalization, three of the four samples were averaged and became the "Dictionary" of the desired CEPV words.

Once completed, the remaining normalized CEPV was compared to each of the numerical pixel matrices. For each word the percent difference was calculated on an elemental row and column basis. The resultant element with the least error was set to zero and all other entries were calculated as a percent error from that result.

Table 1 is the resultant confusion matrix for the comparison. Examination of the confusion matrix shows all zeros along the principal axis. There is a 100% recognition between the unknown word CEPV and the "Dictionary" CEPV. Moreover, the percent error between the unknown and the nearest CEPV is 27 percent whereas most are well into the 100's of percent difference. This result shows the robustness of the sampling and processing technique.

Table 1. Confusion Matrix CEPV Analysis

[[0.	121.	260.	472.	264.	320.	69.	496.]
[27.	0.	228.	529.	205.	292.	63.	558.]
[89.	71.	0.	187.	96.	134.	93.	220.]
[171.	164.	137.	0.	175.	191.	146.	233.]
[220.	131.	184.	444.	0.	88.	196.	418.]
[435.	340.	343.	679.	174.	0.	405.	583.]
[35.	27.	94.	300.	119.	172.	0.	313.]
[306.	292.	249.	338.	263.	252.	290.	0.]]

4. PROPOSED LIP READING SYSTEM

A Covert Advanced Communication for Tactical Urban Security or CACTUS is proposed. The integration of a normal communication system, an eye-safe near infrared laser, a high speed Charge Coupled Device (CCD) with narrow band optical filter with processor could be integrated into a tactical helmet of a unit in a tactical team. This would allow stealth infiltration into an enclosed environment with non-verbal communication without line of sight nor use of hands as in hand signals. This would allow for better awareness of the surroundings of every unit within the tactical team.

4.1 Headset

A cartoon of a tactical helmet is given in Figure 4. Features of the helmet include extensions for the optical source (a) and the CCD camera (b) at a scattering angle near 90 degrees. Included are stereo microphones (c) to enhance hearing within the tactical space. This same direct connection to the user's hearing also protects the member by limiting noise from loud sources such as explosions. A simple chin transducer (d), senses motion to capture the non-verbal CEPV motion to be processed.

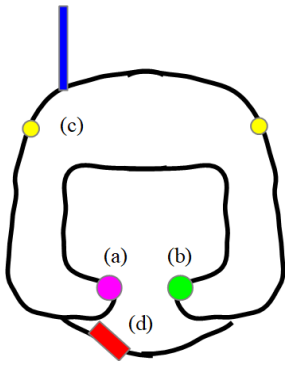


Figure 4. CACTUS Helmet with Features

4.2 Field Programming

Field programming for a particular mission would consist of agreement of the "Dictionary" within the tactical team. Since CEPV's are composed of normal facial expressions the learning curve is minimal. The user would simply place the system into learning mode. A particular word is uttered. Repetition is used to average a number of samples to obtain a representative

CEPV. Upon completion, the system may be tested for robustness by repeating the facial gesture and checking compliance. The standard "Dictionary" may be augmented for specific words for a specific mission, such as the target of a mission. Moreover, a direct translation matrix for a user with a different language can be used. The same words within the tactical "Dictionary" are used in the user's native language and translated to whoever on the tactical team speaks a different language.

4.3 Other Applications

As mentioned, high noise environments, such as flight line operations could benefit from a CACTUS-like system. CACTUS could also be used to enhance audio to text for device control in applications like robotics or Unmanned Ariel Systems. Changes in speech patterns of humans could also be tracked for diagnostic purposes. Finally, surveillance of verbal communications at a distance could be possible if a non-user specific "Dictionary" can be made. A confocal laser detector with proper optics would be required.

The entire method of Coherent Detection of Time Domain Scatter may be applied to a wide array of problems. Any short term time harmonic frequency could be extracted from the scatter modulation of the fundamental optical source frequency. A simulation of extraction of the pulse of a human by the modulation of scatter has been completed and will be published in a future paper. Another example would be monitoring the breathing of a patient within care. It should be possible to not only monitor the respiration rate but also the depth of respiration with changes as a function of time.

5. CONCLUSIONS

The method of Coherent Detection of Time Domain Scattering of light has been presented. The notion of representing an entire word, consisting of an ensemble of phonemes has been given. CACTUS has used the concept of tracking each pixel within the time domain to create a composite representation of a word as a matrix. A "Dictionary" has been used to extract a unique assignment to an unknown word with a 100% success rate. The compression of sampled data to CEPV has been given as 500:1. Over sampling allows for reduced noise properties and language flexibility.

The application to tactical teams has been explored and CACTUS would enable non-verbal communications with high reliability. All components used to create the extraction process are components-off-the-shelf.

5.1 FUTURE WORK

Prototype helmet is in the process of being built. This should allow for much easier expansion of the word sampling. A general processor approach will continue until an optimal sampling rate and modulation rate are settled upon. Expansion into other areas, including biometrics and robotic vision are of current interest.

6. ACKNOWLEDGMENTS

False color image, Fig. 4., has been used with the permission of the participant. This work and all results are the sole ownership of the author. No resources of any current or former employer or affiliation have been used in the concept, development or creation.

7. REFERENCES

[1] S.K. Gaikwad, B.W. Gawali, Pravin Yannawar, "A Review on Speech Recognition Technique", International Jour. of Computer Applications, Vol. 10 - No. 3, November 2010, pp16-24.

[2] M. Battista, F. Collesei, E. Orzan, M. Fantoni and D. Bottari, "Lip-Reading: Advances and Unresolved Questions in a Key Communication Skill, Adiol. Res., July 21, 2025.

[3] C. C. Jorgensen and D. D. Lee, Ames Research Center, NASA Tech Brief, May 1, 2009, pp. 37-38.

[4] Gilbert JM, Rybchenko SI, Hofe R, Ell SR, Fagan MJ, Moore RK, Green P. Isolated word recognition of silent speech using magnetic implants and sensors. Med Eng Phys. 2010 Dec; 32(10):1189-97.

[5] Y. Assael, B. Shillingford, S. Whiteson and N. de Fretias, "LipNet: End-to-End Sentence-Level Lipreading", arXiv:1611.01599, 2016.

[6] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, “CUAVE: a new audio-visual database for multimodal human computer- interface research”, Proceedings of IEEE. International conference on Acoustics, speech and Signal Processing, 2017-2020, 2002

[7] A. Bizzocchi, "How Many Phonemes Does the English Language Have?", Int. Jour. on Studies in English Lang. and Lit., Vol. 5, Issue 10, October 2017, pp. 36-46.

[8] B. Kanack, "Measurement of Intermodulation Distortion in Optical Diodes", 1995 IEEE MTT-S Digest, pp. 61-64.