Can lip contours help us improve understanding of speech in noisy environments?

Arun Palghat Udayashankar[1,2], Meghan Stansell[1,2], Gabrielle Saunders[1,2] & Peter G. Jacobs[1,2]

1.National Center for Rehabilitative Auditory Research (NCRAR), Portland VA Medical Center
2.Oregon Health and Sciences University

Summary

Showing a speakers face is known to improve speech recognition in noise performance compared with the audio-alone condition. Here, we explore how modulation frequency envelopes correlate with lip motion. Modulation filtering of speech involves decomposing the signal into frequency bands and then dividing each band into a high frequency carrier and a low frequency modulation envelope. The modulation envelope temporally shapes the carrier frequency in a way that is similar to the way the lips shape sound energy during speech production. We found a good correlation between lip contour movements and modulation envelopes as recorded during audio-visual speech at the sentence, word, and simple-sound level. Our group and others have shown a marked improvement in voice activity detection by adding visual cues particularly under poor SNR conditions (< 0 dB SPL). These results encouraged us to adapt an algorithm previously proposed by Kim et al. (2009), to use a multi-band binary MASK approach to improve intelligibility of speech under noisy conditions using both audio and visual features. Here we present a classifier that uses the same audio features described in the above study and integrate that with video features for classification. This results in a marked improvement in classification and sound quality. In the future we would like to generate sound files with the proposed algorithm to clean up noisy speech files and perform intelligibility tests on normal listeners and patients with dual sensory hearing loss to quantify the benefit of the proposed algorithm towards speech intelligibility. If shown to be beneficial, we propose to implement the algorithm in an audio-visual hearing aid.

.

Research questions

- **Broad Question:** Can visual features (i.e. lip and face movement) be used to improve speech quality in a hearing aid under noisy conditions?
- **Today's Question:** How can visual features be combined with auditory features to improve classification accuracy in a signal processing algorithm that filters out noise using a spectrotemporal binary mask?

Proposed audio-visual hearing aid system

Hearing aid receives visual features and runs signal processing algorithm that fuses audio and visual information to improve acoustical SNR

Miniature video camera captures face movement and extracts relevant mouth features.

Approach

- Record audio and visual corpus of two speakers reciting 400 revised speech-in noise (R-SPIN) sentences.
- Add different levels of broadband white noise to the acoustic signals at SNRs of 0 dB, -6 dB, and –12 dB.
- Train the proposed classifier using audio and visual cues on 90% of the recorded sentences.
- Test the classifier on the remaining 10% of the data and determine classifier accuracy.
- Evaluate accuracy using acoustic features alone, visual features alone, and audio + visual features on individual sentences.

Modulation envelopes and correlation with lip motion

Any given speech signal can be filtered into sub bands. Each sub-band signal in turn can be decomposed into a high frequency carrier and a low frequency envelope that modulates the carrier. This is illustrated in Figure 1. The red curves represent carriers and the blue curves are examples of modulators (figure reproduced from Steven Schimmel 2007). In the present study, in addition to the 400 sentences from the RSPIN corpus, we recorded simple sounds along with the corresponding video of the speaker talking. The audio signal was filtered into sub-bands with bandwidth of 500 Hz. We used the Luxand FaceSDK face recognition software to extract 66 points on the face of the subject including lip height and lip width for each video frame. We correlated the lip height (separation between upper and lower lip at the center) with the modulator of the first band in the case of sentences, words (Figure 2) and simple sounds (Figure 3). The infographics show words and simple sounds with font sizes scaled according to the R^2 values of the correlation. Only a selection of words that appeared more than twice in the 400 sentences and had an $R^2 > 0.45$ are shown. The correlation for words varied between 0.2 to 0.65. Around 92 simple sounds are shown below and the corresponding R^2 values varied between 0.4 and 0.9. In the case of full sentences, the correlations varied between 0.45 to 0.76. Despite the high values of correlations, it was found that swapping out the modulators with trajectories of variation of lip height for the sentences did not retain speech intelligibility.

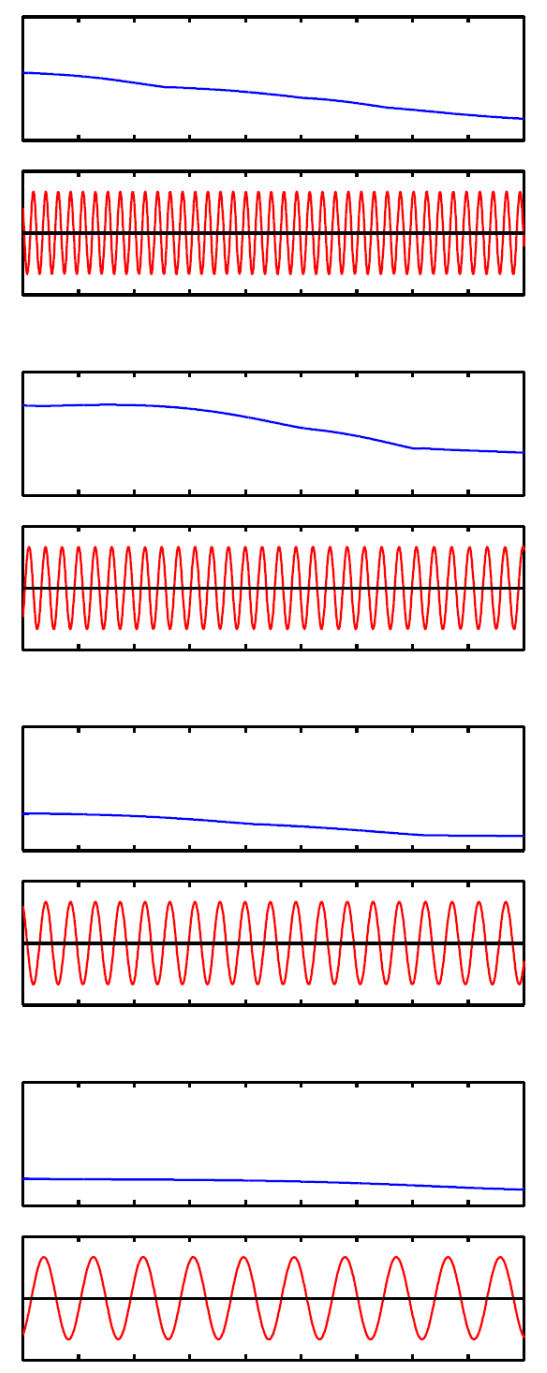


Figure 1

Words

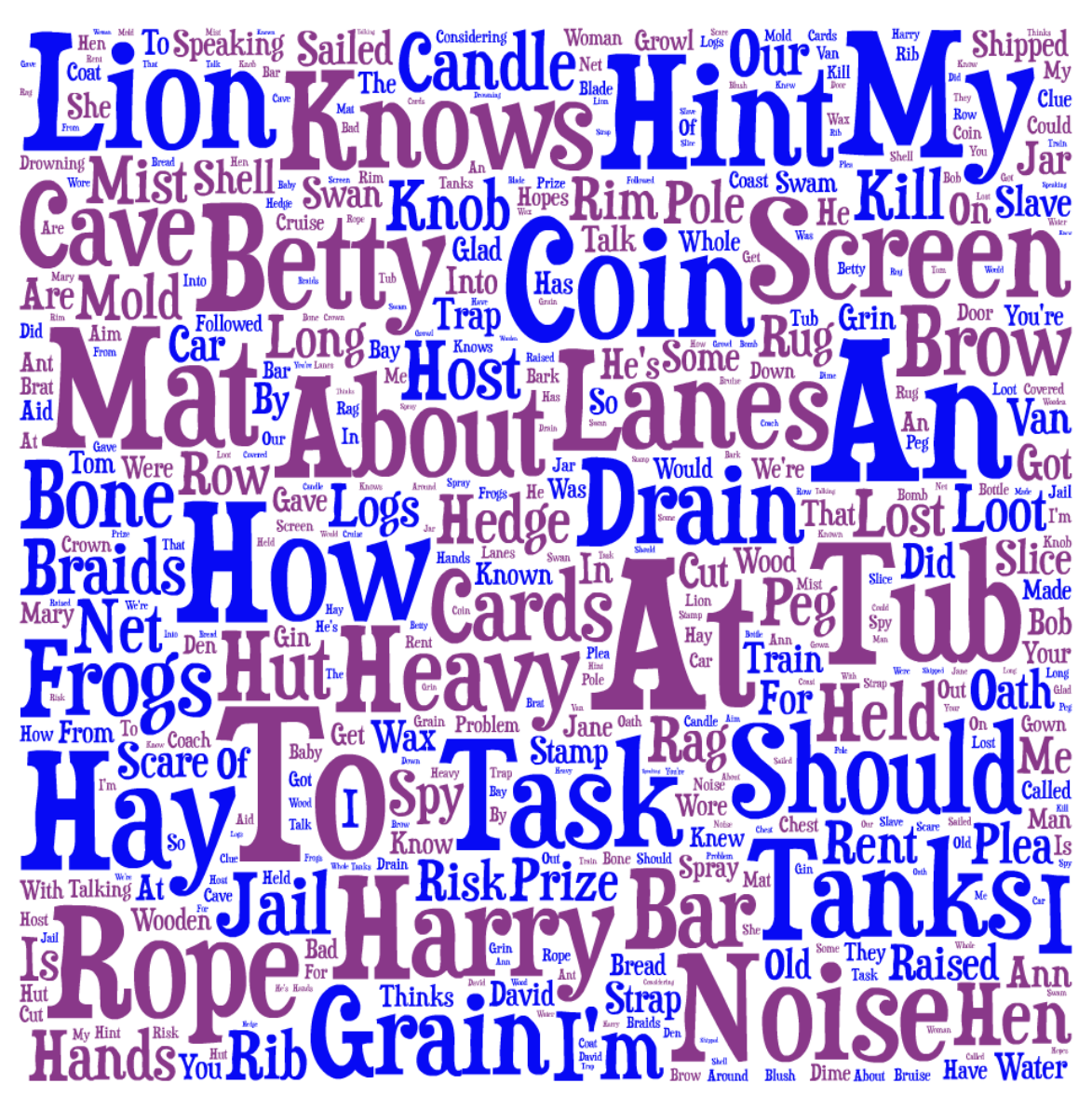


Figure 2

Simple sounds

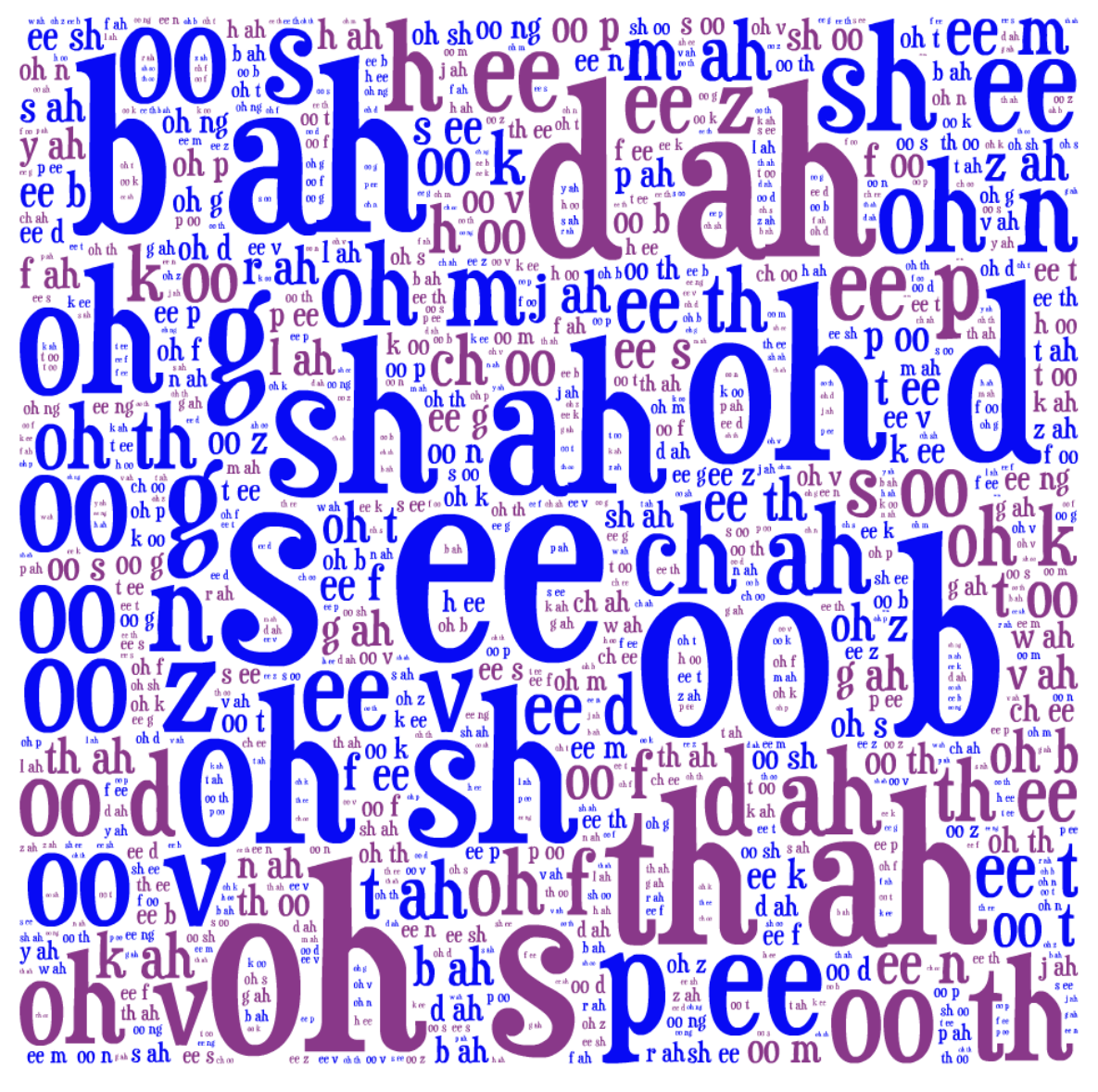
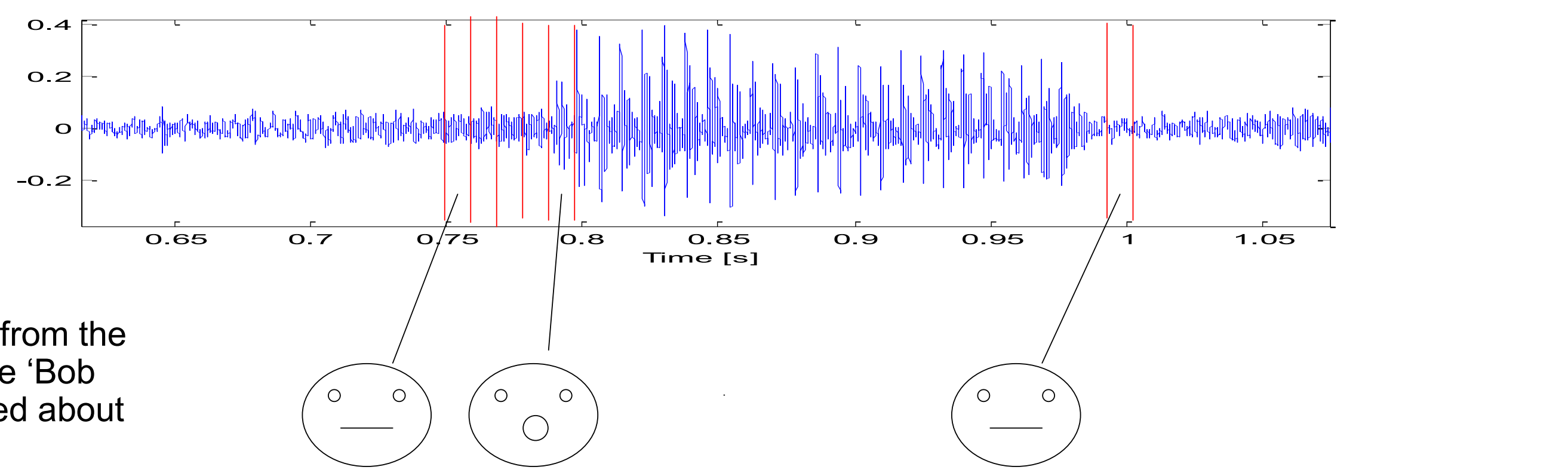


Figure 3

Carriers (red) and modulators (blue)

Audio and visual feature extraction

Audio segmented into 10.4 ms frames with 40% overlap

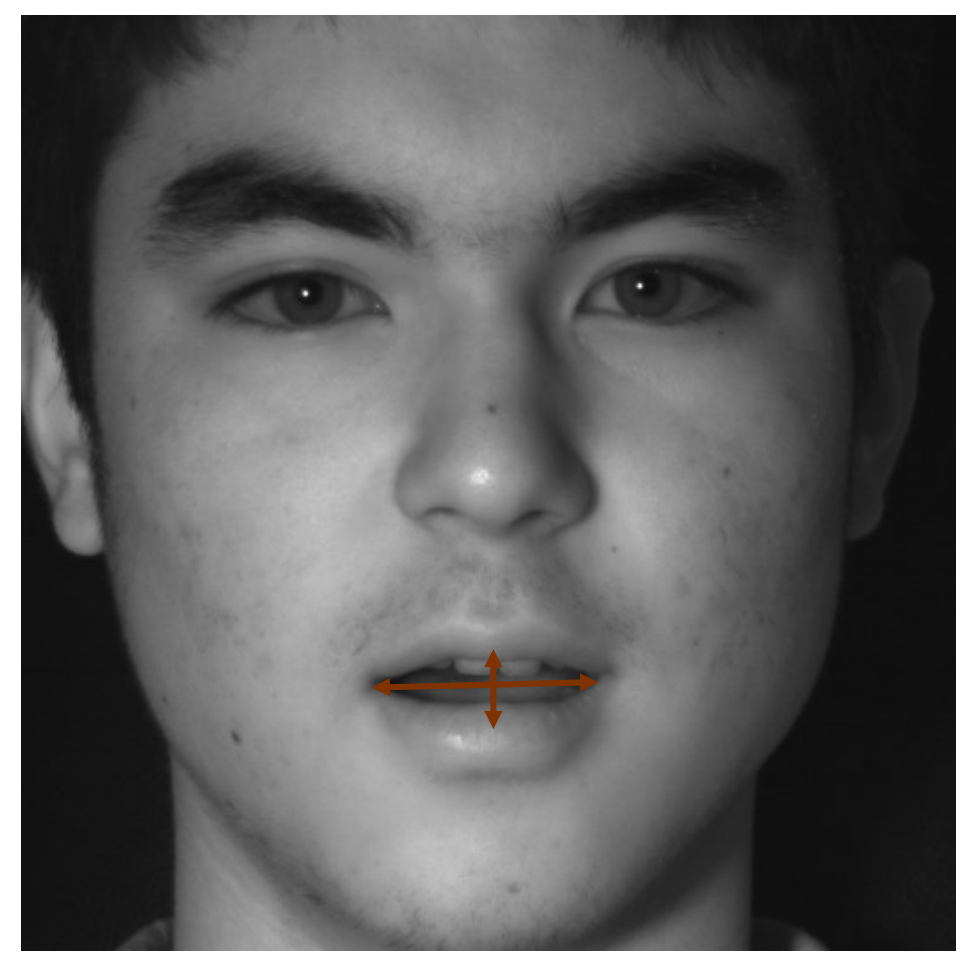


The word 'Bob' from the RSPIN sentence 'Bob heard Paul called about the strips'

Video frames (96 fps) aligned with audio segments (44,100 samples/sec)

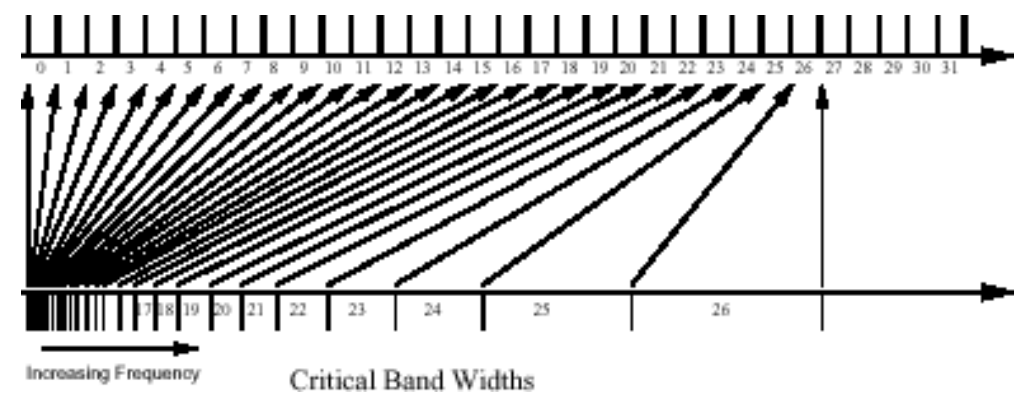
Visual Feature Extraction

- Visual Features extracted using 66 markers around the face
- From these markers, the following features were extracted:
  1. Lip height upper lip to lower lip
  2. Lip width
  3. Area of ellipse fit to the above two
  4. Δ height, Δ width & Δ area
  5. Total of 6 visual features were used
-



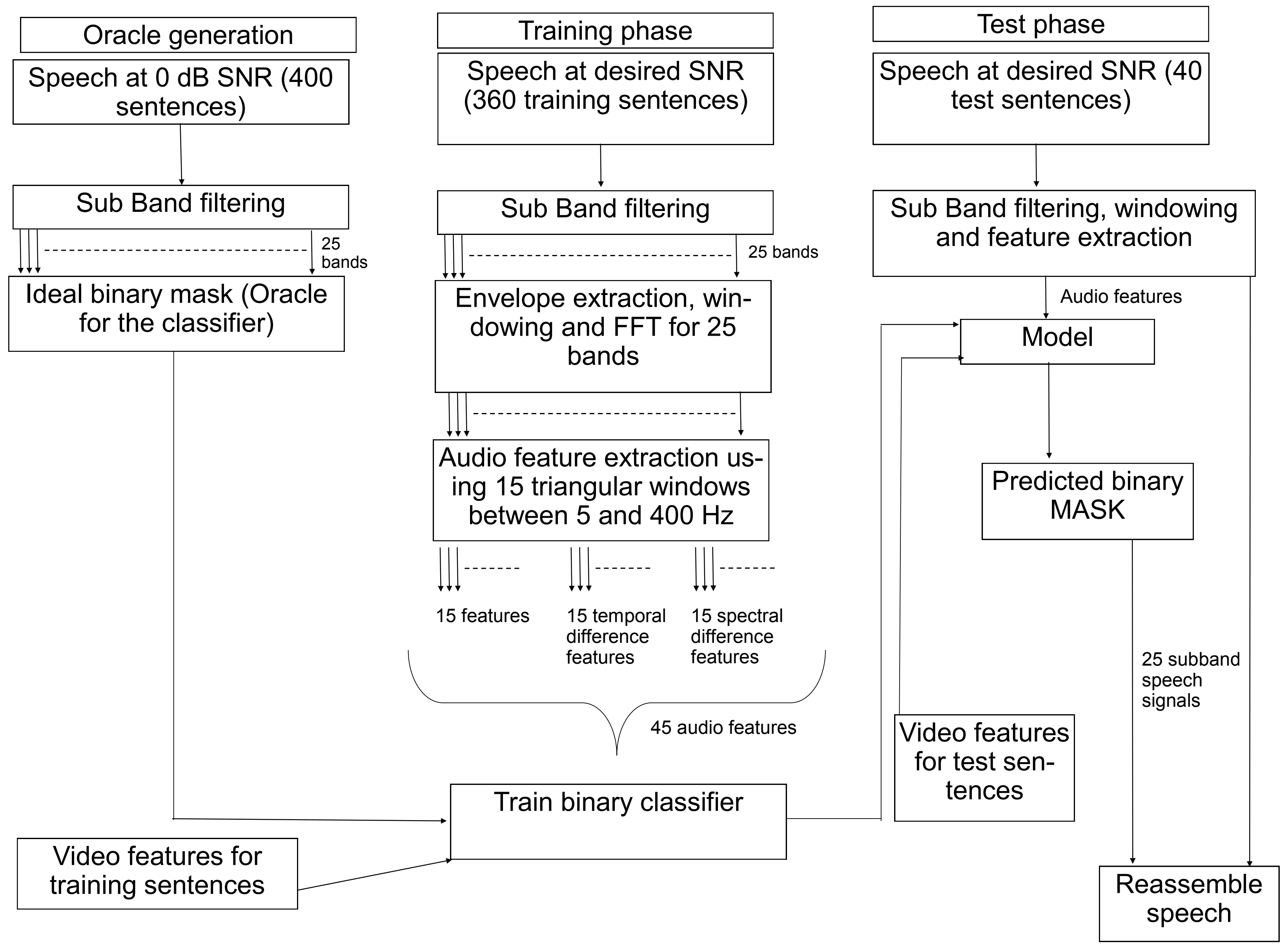
Auditory Feature Extraction

Sound was sub-band filtered into the first 25 critical bands of human hearing. The filtered signal for each band was windowed using a Hamming window of width ~11 ms. and the FFT of the envelope signal for each window was computed. Using 15 equally spaced triangular filters with no overlap in the frequency range 5-400 Hz, the auditory features were extracted. The features were a function of time and frequency, denoted as a(t,k). Taking the difference between adjacent time windows gave $\Delta a_t(t,k)$ and adjacent frequency bands gave $\Delta a_k(t,k)$. This led to a total of 45 (15*3) auditory features (Kim et al. 2009).



Proposed signal processing algorithm

The proposed signal processing algorithm (modified after Kim et al. 2009) is shown below. The recorded sound files are noised at the required SNR level with broadband white noise. Audio features are extracted after sub-band filtering and envelope extraction by taking the magnitude of the Hilbert transform. This is followed by windowing with a Hanning window for a duration of 11ms with 50% overlap. This leads to dividing the incoming speech signal into frames. The Signal to noise ratio is then calculated by taking a ratio between the incoming speech signal and the noise for any given time window. If the SNR is found to be above a certain threshold, it is classified as signal and vice versa if the SNR is found to be below the threshold. In this manner, each time window in the various bands are populated with 1s (signal) and 0s (noise) forming the the binary MASK. An ideal binary MASK is made at a good SNR level (> 0dB SPL) and used as the Oracle for the classifier. A simple regression model was fit in this case for each frequency band (25) to all the 51 features (45 audio and 6 video). The model was pruned to remove any features that gave a poor fit ($p<0.05$). Then the features for any given band were standardized across all sentences by subtracting the features from the mean and dividing by the standard deviation (μ and σ). The μ and σ values were saved to condition features in the enhancement stage. In the enhancement stage, the incoming audio signal was subband filtered, windowed and the features were extracted. The features were conditioned with the μ and σ from above. Based on the features, for any band and window the classification model was applied to generate the binary MASK. The subband filtered and windowed signal was then multiplied with the predicted binary MASK and all the resulting subband signals were added up together to synthesize the speech signal.



Oracle generation

Speech at 0 dB SNR (400 sentences)

Sub Band filtering

25 bands

Ideal binary mask (Oracle for the classifier)

Video features for training sentences

Training phase

Speech at desired SNR (360 training sentences)

Sub Band filtering

25 bands

Envelope extraction, windowing and FFT for 25 bands

Audio feature extraction using 15 triangular windows between 5 and 400 Hz

15 features

15 temporal difference features

15 spectral difference features

45 audio features

Train binary classifier

Test phase

Speech at desired SNR (40 test sentences)

Sub Band filtering, windowing and feature extraction

Audio features

Model

Predicted binary MASK

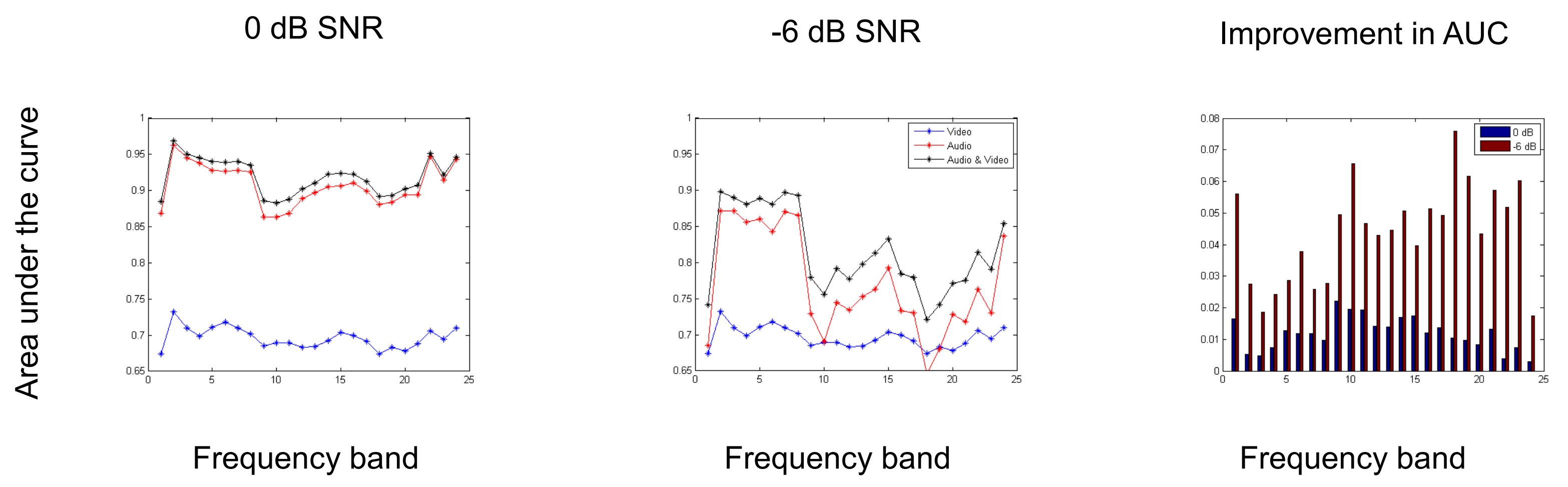
25 subband speech signals

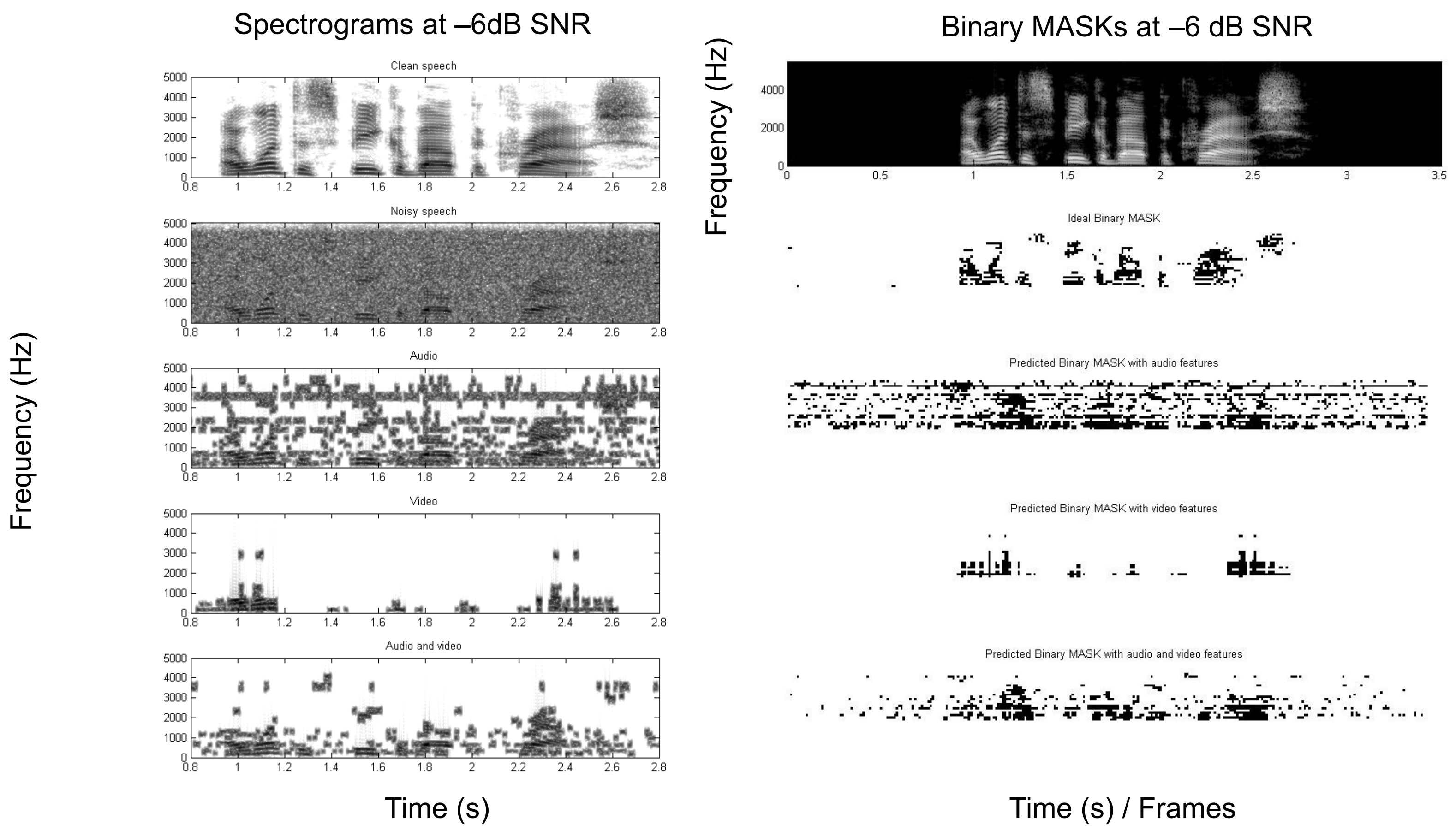
Video features for test sentences

Reassemble speech

Accuracy of classification based on audio and visual cues

A previous study (Kim et al. 2009 &) showed great success in improving speech intelligibility using the ideal binary MASK approach and a Bayesian classifier even at –5 dB SNR for different types of noise. Our results show that a simple regression based classifier works well at 0 dB SNR. Adding the video features did improve the classification rate by a small amount. However as the SNR is lowered to –6 dB, the benefit of adding the video features to the classifier is substantially higher as quantified by the improvement in the area under the curve. This benefit is further illustrated by means of an example at –6 dB SNR. The bottom right panel shows spectrograms of the original, noised, classified audio files for the "video", "audio" and "audio+video" conditions. It can be seen that the video alone condition does a good job of being a voice activity detector and gets rid of a lot of high frequency noise. The audio alone condition reproduces the signal but fails to get rid of the noise in some of the higher frequency bands due to misclassification. Adding the video features to the audio helps improve the classification by retaining most of the signal and getting rid of a lot of broadband noise that the audio alone condition couldn't filter out. The bottom left panel shows the binary MASKs for the different conditions below the original audio file including the oracle for the classifier (ideal binary MASK). In conclusion, adding the video features helped us extend the benefit of the regression based audio classifier to –6 dB SPL (see spectrograms and binary MASKs). Going into the future, we would like to push this limit further down to lower SNRs which would be potentially benefit people with hearing loss who have difficulty even at relatively high SNR conditions like 0 dB SPL.





Conclusions and future directions

- Visual cues including lip height and width improve classification accuracy in the proposed algorithm.
- Benefit of visual cues for classification accuracy increases as the SNR is lowered (< 0 dB).
- Audio cues dominate in high SNR conditions while visual cues dominate in low SNR conditions.
- The ideal binary mask approach has been shown previously as being effective for improving speech intelligibility at –5 dB SNR. This limit may be lowered further by integrating visual cues.
- Visual feature extraction speed needs to be improved (currently 100 ms to extract 66 features). Since all 66 points are not required, speed may be improved by focusing solely on the mouth.
- In the future, we will test our model on data recorded from other speakers.
- In the future, we will experiment with lower SNR conditions (< -6 dB SPL).

References

Peter G. Jacobs, Deniz Erdogmus, Sarah Weidman, Gabrielle Saunders (2012) Methods on improving signal-to-noise ratio of a speech signal using visual features from a speaker's face, International Hearing aid Conference

Steven M. Schimmel (2007) "Theory of Modulation Frequency Analysis and Modulation Filtering, with Applications to Hearing Devices," Ph.D. Dissertation, University of Washington

Kim, G., Lu, Y., Hu, Y. and Loizou, P. (2009). "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," Journal of Acoustical Society of America, 126(3), 1486-1494

Contact: *palghatu@ohsu.edu* & *jacobsp@ohsu.edu*

Project support came from the Department of Veterans Affairs RR&D C7693M and Oregon Health and Science University