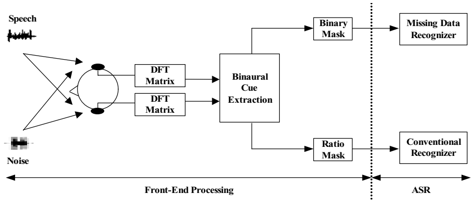
ON BINARY AND RATIO TIME-FREQUENCY MASKS FOR ROBUST SPEECH RECOGNITION

Soundararajan Srinivasan, Nicoleta Roman and *DeLiang Wang*, The Ohio State University, {srinivso, niki, dwang}@cse.ohio-state.edu

SUMMARY

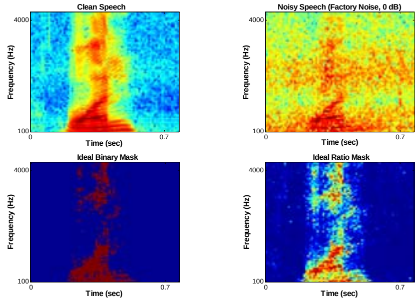
A time-varying Wiener filter specifies the ratio of target signal and a noisy mixture in a local time-frequency unit. We estimate this ratio using a binaural processor and derive a ratio time-frequency mask. This mask is used to extract the speech signal, which is then fed to a conventional speech recognizer operating in the cepstral domain. We compare the performance of this system with a missing data recognizer that operates in the spectral domain using the time-frequency units that are dominated by speech. To apply the missing data recognizer, the same binaural processor is used to estimate an ideal binary time-frequency mask, which selects a local time-frequency unit if the speech signal within the unit is stronger than the interference. We find that the performance of the missing data recognizer is better on a small vocabulary recognition task but the performance of the conventional recognizer is substantially better when the vocabulary size is larger.

1. SYSTEM ARCHITECTURE



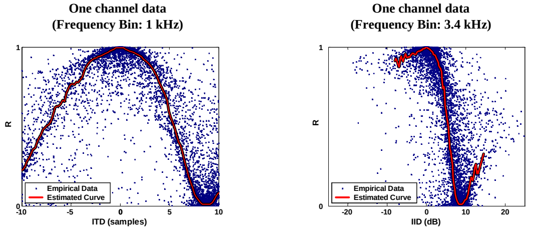
- The binaural input is obtained by convolving the monaural signals with measured head related impulse responses (HRIR) from a KEMAR dummy head. A short-time Fourier analysis is then used to derive a time-frequency (T-F) decomposition. Interaural time differences (ITD) and interaural intensity differences (IID) are computed in each T-F unit.

2. IDEAL BINARY AND RATIO MASKS



- An ideal binary mask is defined *a priori* by selecting the T-F units where speech energy is stronger than interference energy.
- An ideal ratio mask is defined *a priori* as the ratio of speech energy to total energy (speech and noise) in each T-F unit.
- Binaural-based methods are used for the estimation of the two masks.
- The missing-data recognizer (Cooke et al., 2001) attempts to achieve robust speech recognition by distinguishing between reliable and unreliable data in the T-F domain. Hence, the binary mask is used as input to this recognizer.
- The speech signal resynthesized using the ratio mask is used as input to the conventional ASR.

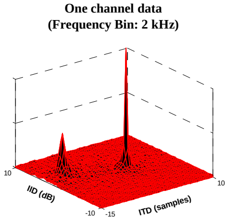
3. RATIO MASK ESTIMATION



Data scatter plots for ITD/IID and the *a priori* energy ratio *R*.

Statistics for the relationship between the *a priori* energy ratio *R* and the pattern of binaural cues are at the core of our binaural processor. We employ a training set containing 10 speech utterances from the TIMIT database. The ITD/IID are computed in each T-F unit based on the spectral ratio at the left and right ears. We show for mixtures of multiple sound sources, there exists a strong correlation between the energy ratio and the ITD/IID values. The estimated curve is the empirical mean. Given the ITD/IID values in a local T-F unit, the estimated energy ratio is computed as the corresponding value on the mean curve.

4. BINARY MASK ESTIMATION



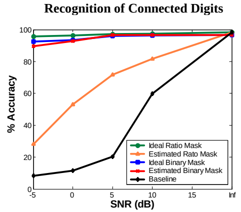
Histogram for ITD/IID

Given a spatial configuration, the ITD/IID statistics show a characteristic clustering across frequency bins. Each peak in the histogram corresponds to a distinct active source. The estimated binary mask is obtained using non-parametric classification in the joint ITD-IID space as used by Roman et al. (2003).

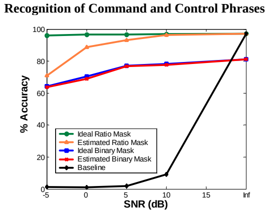
5. RESULTS

Evaluation Setup

- The connected digits recognition task (perplexity 11.0) uses the male speaker dataset from the TIDigits database. The command and control task (perplexity 3.05) uses the digital dataset of the Apple Words and Phrases database.
- The noise source is factory noise from the NOISEX database.
- For both tasks, the target source is in median plane and noise source on the right side at 30º.
- Word level HMMs are trained for the two recognizers. Each HMM comprises of 8 emitting states whose output probability is modeled using Gaussian mixtures with diagonal covariance.
- While the missing-data recognizer operates in the log-spectral domain, the conventional ASR operates in the cepstral domain.
- "Baseline" refers to the performance of the conventional ASR without any front-end processing.



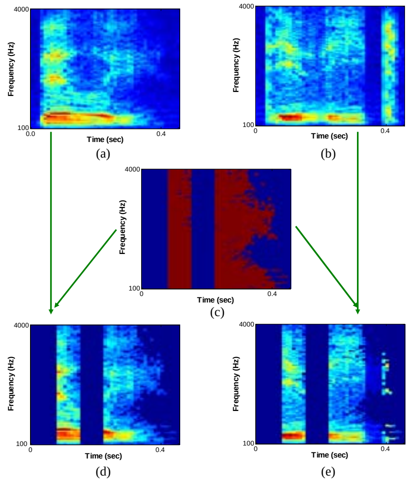
- Conventional ASR shows only a minor degradation with ideal ratio mask. Its performance with estimated ratio mask degrades much faster (but significantly better than baseline).
- The missing-data recognizer with ideal binary mask shows very little degradation and its performance with estimated binary mask is close to that with ideal binary mask, indicating the front-end's ability to estimate ideal binary mask accurately.



- Conventional ASR with ideal ratio mask achieves close to ceiling performance. Its performance with estimated ratio mask is close to that with ideal ratio mask (especially at SNR > 0 dB) and substantially better than the performance of the missing data recognizer with both ideal and estimated binary masks.
- Lower accuracy of the missing data recognizer under clean speech may be due to diagonal covariance assumption.

6. EFFECT OF BINARY T-F MASK ON SPECTRUM

This is an illustration of the similarity between the reliable spectral regions of two words: (a) "Billy" and (b) "Delete". The outcome of applying the binary mask in (c) to the spectrogram of these two words is seen in (d) and (e). The resulting reliable regions of the two spectrograms are very similar. In the absence of information in the unreliable regions, it is difficult for the missing-data recognizer to distinguish between the two words.



7. CONCLUSION

- **For the tasks considered in the present study, we have proposed a method to estimate a ratio T-F mask front-end for a conventional ASR, using a binaural processor and demonstrated substantial improvement over the baseline performance over a range of SNRs.**
- **On a lower perplexity task, the performance of the conventional ASR using the estimated ratio mask is substantially better than that of the missing data recognizer.**
- **The degradation in the performance of the missing-data recognizer on the larger vocabulary task may be caused by its inability to represent a larger number of speech models on a common T-F representation adequately.**