Can binary masks improve intelligibility?

Mike Brookes (Imperial College London) &
Mark Huckvale (University College London)

Apparently so ...

An algorithm that improves speech intelligibility in noise for normal-hearing listeners

Gibak Kim, Yang Lu, Yi Hu, and Philipos C. Loizou[a)]
Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas 75080

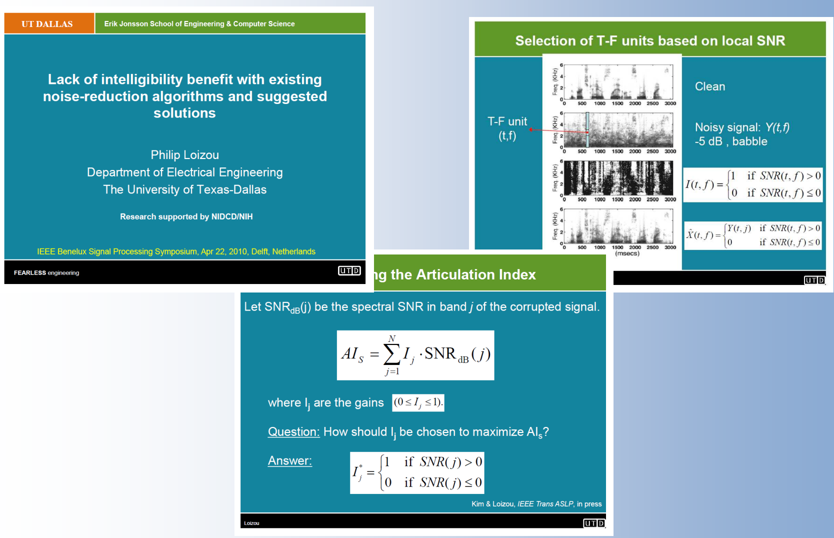
(Received 30 October 2008; revised 27 March 2009; accepted 1 July 2009)

Traditional noise-suppression algorithms have been shown to improve speech quality, but not speech intelligibility. Motivated by prior intelligibility studies of speech synthesized using the ideal binary mask, an algorithm is proposed that decomposes the input signal into time-frequency (T-F) units and makes binary decisions, based on a Bayesian classifier, as to whether each T-F unit is dominated by the target or the masker. Speech corrupted at low signal-to-noise ratio (SNR) levels (−5 and 0 dB) using different types of maskers is synthesized by this algorithm and presented to normal-hearing listeners for identification. Results indicated substantial improvements in intelligibility (over 60% points in −5 dB babble) over that attained by human listeners with unprocessed stimuli. The findings from this study suggest that algorithms that can estimate reliably the SNR in each T-F unit can improve speech intelligibility.
 [DOI: 10.1121/1.3184603]

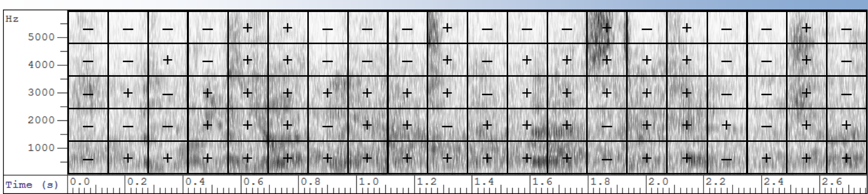
PACS number(s): 43.72.Ar, 43.72.Dv [MSS] Pages: 1486–1494

J. Acoust. Soc. Am. **126** (3), September 2009

How does it work?



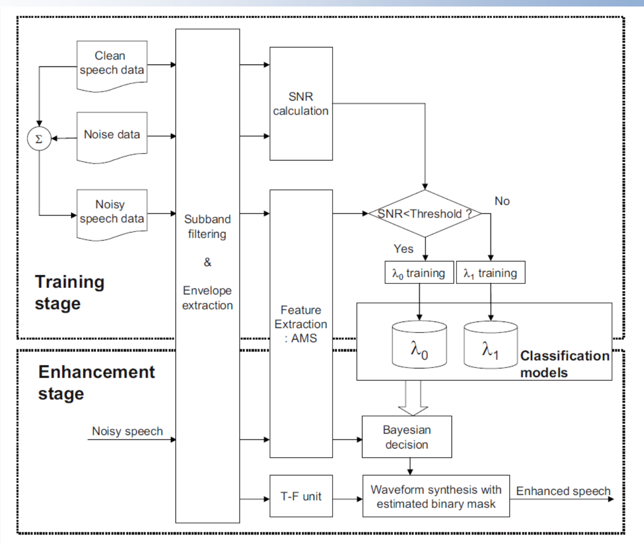
Time-frequency grid of local SNR



$$intelligibility\ index = F\left(\sum_f w(f)\left\{\frac{\sum_t e_s(t,f)}{\sum_t e_n(t,f)}\right\}\right)$$

- e_s = speech energy, e_n = noise energy, w() = frequency weighting
- F() is some monotonic function
- index is increased if attenuation applied in each cell where $e_n > e_s$
- i.e. where local SNR < 0dB

Use of classifier to estimate binary mask



Replication

Similarities

- IEEE sentences as training testing materials
- Single male talker
- Babble and speech-shaped noise @ -5dB SNR
- Signals at 12,000 samples/sec
- Acoustic features based on modulation spectrum - code provided by Kim
- Feature vector incorporates time & frequency deltas
- SNR thresholds for constructing target mask on training data
- GMM classifier design, using full covariance
- Four GMMs to classify feature vectors based on division of training vectors into groups based on SNR.

Differences

- We used a different, British English, talker
- We used babble from NOISEX ROM

Thanks to: Toby Davies

Classifier performance (@ -5dB SNR)

SNR > 0 Cells %	Speech-shaped noise		Babble noise	
	Hits	False-Alarms	Hits	False-Alarms
Kim et al	88.3	9.5	87.0	14.5
Ours				

Classifier performance (@ -5dB SNR)

SNR > 0 Cells %	Speech-shaped noise		Babble noise	
	Hits	False-Alarms	Hits	False-Alarms
Kim et al	88.3	9.5	87.0	14.5
Ours	55.2 ☹	15.0	51.6 ☹	15.0

Intelligibility performance (@ -5dB SNR)

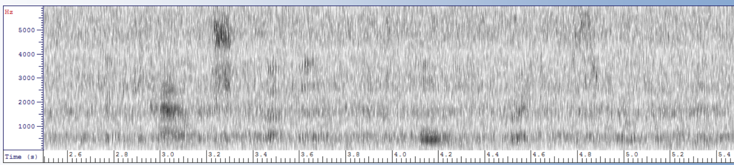
Words %	Speech-shaped noise			Babble noise		
	No proc.	Proc.	Ideal	No proc.	Proc.	Ideal
Kim et al	45	87	92	19	85	92
Ours						

Intelligibility performance (@ -5dB SNR)

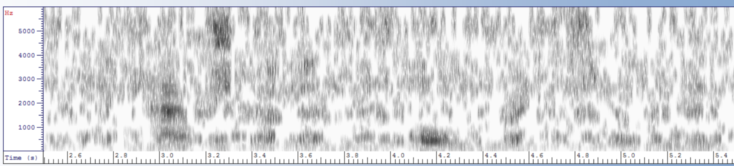
Words %	Speech-shaped noise			Babble noise		
	No proc.	Proc.	Ideal	No proc.	Proc.	Ideal
Kim et al	45	87	92	19	85	92
Ours	49	21 ☹	77	54	15 ☹	85

Binary Mask Enhancement

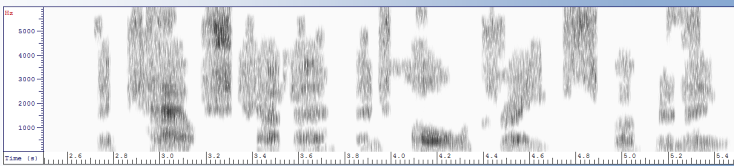
LTASS
-5dB



Recognised
Mask

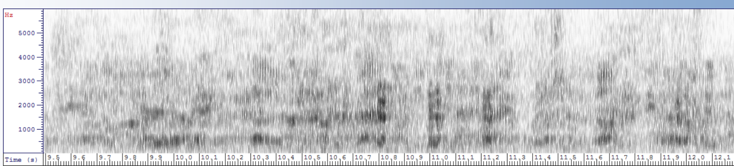


Ideal
Mask

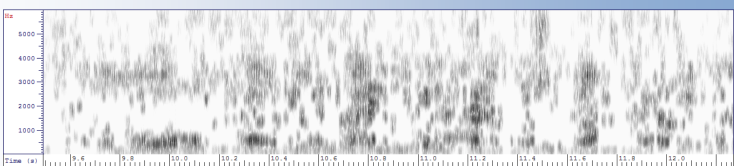


Binary Mask Enhancement

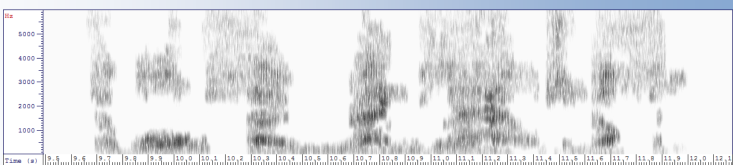
Babble
-5dB



Recognised
Mask



Ideal
Mask



What is going on?

- There are a number of arbitrary parameter settings in Kim et al (2009)
  - Sampling rate, window size, number of channels
  - Down-sampling of modulation spectrum
  - SNR thresholds for binary mask choice
- These may have become “optimised” for particular data set they used
- Overall performance may be very sensitive to small changes in system design
- We need to investigate and understand details of algorithm
- ... over to Mike

What is the perfect binary mask?

- **Original idea [Wang2005]:**
  Select Time-Frequency (TF) cells with $S(t, f) - N(t, f) > L$
  where S and N are speech and noise power spectral densities in dB and L is a threshold (“Local Criterion”)

- **Motivation: Masking**
  Exclude TF cells with poor SNR since they give little information and may mask adjacent frequency bands

- **However …**
  If we plot intelligibility versus L for different SNR levels the results do not match this theory

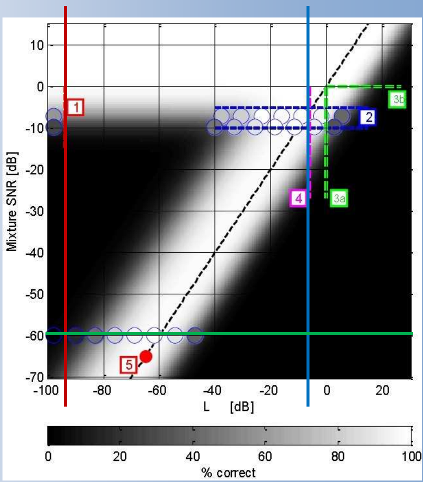
Intelligibility of Binary Masked Speech

- $L=-\infty$:OK@ > –10dB SNR
- $L=-6$:OK@ > –20dB SNR
- SNR=–60 dB:
  OK @ –90 dB < L <-50 dB

Two independent sources of information [Kjems et al 2010]:

1. Noisy speech signal
   SNR > –10 & (L–SNR) < 10
2. Noise-vocoded signal
   –30 dB < (L–SNR) < 10 dB

The benefit of binary masking comes entirely from component 2



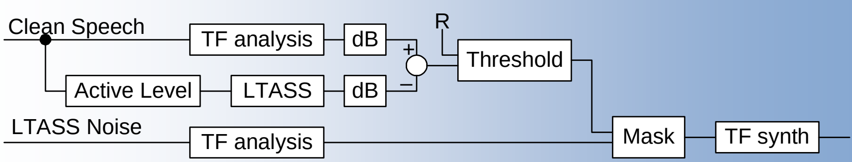
[Kjems et al, EUSIPCO-2010]

Noise-Vocoded component

- Define "Relative Criterion": $R = L - SNR = L - \left(\bar{S}(f) - \bar{N}(f)\right)$
- Mask becomes: $S(t,f) - N(t,f) > R + \bar{S}(f) - \bar{N}(f)$
- Eliminate noise dependency by taking $N(t,f) = \bar{N}(f)$

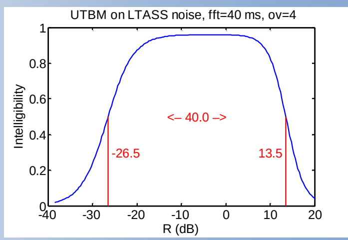
$$S(t,f) - \bar{S}(f) > R$$

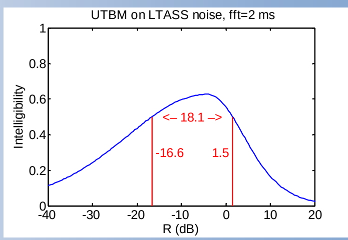
"Target Binary Mask"



Unimodal Psychometric Function Modelling

- Product of two logistic curves
  - Fixed guess/lapse rates
  - 4 free parameters
- Modify to remove interaction between low and high slopes
  - No change if low and high slopes are equal
  - Negligible change if slopes are widely separated
  - Estimation is easier and more stable
- Use width @ 50% as a single figure of merit
  - Not always ideal

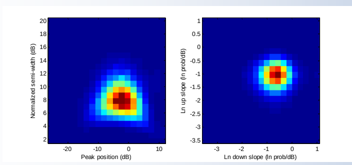


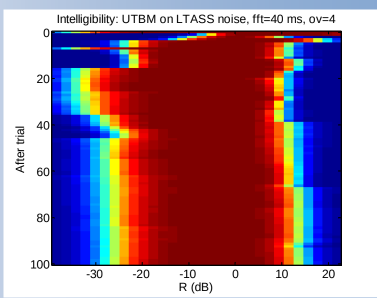


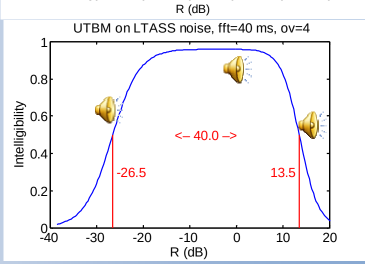
Centre for Law Enforcement Audio Research 17

Psychometric Function Evaluation

- Digit triples: male+female
  - Forced choice experiment
- Bayesian estimation of pdf of 4-D parameter vector
  - Update pdf after each trial
  - Select next R to give greatest expected entropy reduction
  - Very quick convergence (e.g. 60 trials)



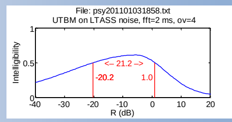




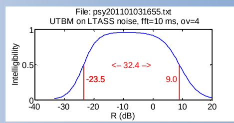
Centre for Law Enforcement Audio Research 18

Effect of FFT length

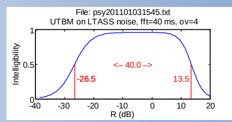
- TF analysis/synthesis
  - Hamming window of length T
  - Freq resolution ~ 1.8/T
  - Modulation bandwidth ~ 0.9/T
- Observations
  - @ T=40ms, R can vary by 40 dB
  - @ T=160ms performance worse: too much smoothing in modulation domain?
  - @ T=2ms performance worse: cannot resolve formants?
  - @ T=10ms performance still OK



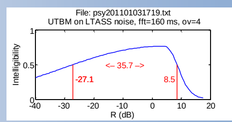
T=2 ms
f_{res}=900 Hz
f_{mod}<450 Hz



T=10 ms
f_{res}=180 Hz
f_{mod}<90 Hz



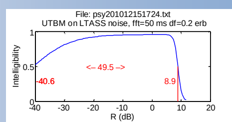
T=40 ms
f_{res}=45 Hz
f_{mod}<22.5 Hz



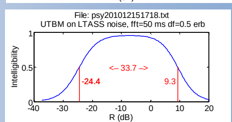
T=160 ms
f_{res}=11 Hz
f_{mod}<5.6 Hz

Non-uniform frequency resolution

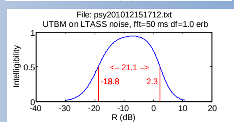
- FFT length kept at 50 ms
  - f_{res}=36 Hz, f_{mod}<18 Hz
- Change mask resolution
  - Estimate mask in erb domain
  - 0.2, 0.5, 1 and 2 erb resolution
- Observations
  - Even at a resolution of 0.5 erb, the intelligibility is noticeably worse [surprising]
  - Substantial degradation at 1 erb resolution



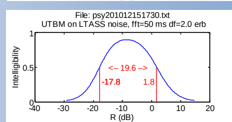
T=50 ms
f_{res}=0.2 erb



T=50 ms
f_{res}=0.5 erb



T=50 ms
f_{res}=1.0 erb

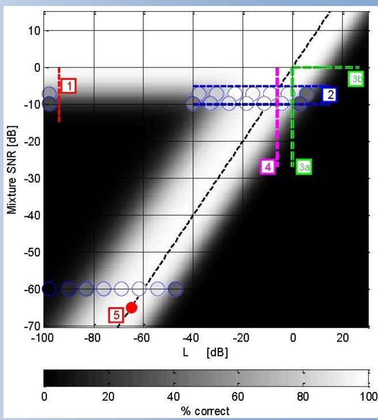


T=50 ms
f_{res}=2.0 erb

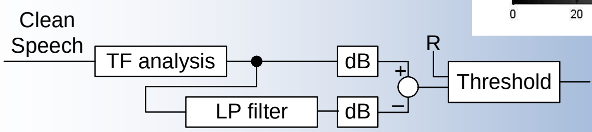
Modulation Domain Model

Intelligibility determined by accuracy of modulation domain spectrum
[Taal et al, ICASSP 2010]

- Encompasses both regions of the graph within one concept
- Measure by correlation coefficient between clean and masked speech in 400ms window for each frequency bin



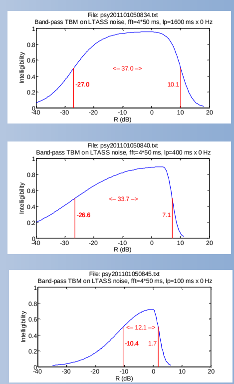
Maximize by comparing with low pass filtered version of spectrogram:



Centre for Law Enforcement Audio Research 21

Time Correlation based mask

- LP filter operates on power spectrum in time domain
  - Hamming window impulse resp
  - LP cutoff = $0.9/T_{LP}$
- Correlation coeff between clean and masked max when R=0
- Observations
  - Poor intelligibility compared to previous for short T_{LP}
  - Very noisy: mask tries to match noise when no speech energy
    - Use noise floor threshold



T_{LP}=1600 ms
F_{mod}>0.6 Hz

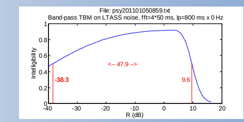
T_{LP}=400 ms
F_{mod}>2.3 Hz

T_{LP}=100 ms
F_{mod}>9 Hz

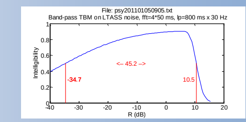
Centre for Law Enforcement Audio Research 22

Time-Freq Correlation based Mask

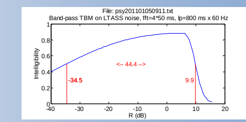
- Seems reasonable to try matching modulation in both time and frequency
  - Apply LP filter in both directions
  - Fix T_{LP}=800 ms giving mod domain HP at 1.1 Hz
  - Vary filter width in frequency direction
- Observations
  - Makes rather little difference
  - F_{LP}=120Hz gives some benefit



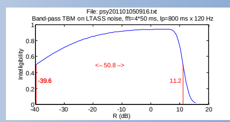
T_{LP}=800 ms
F_{mod}>1.1 Hz
F_{LP}=0 Hz



T_{LP}=800 ms
F_{mod}>1.1 Hz
F_{LP}=30 Hz



T_{LP}=800 ms
F_{mod}>1.1 Hz
F_{LP}=60 Hz



T_{LP}=800 ms
F_{mod}>1.1 Hz
F_{LP}=120 Hz

Summary

- Intelligibility benefits arise from the noise vocoded component of the masked speech
- Rapid estimation of unimodal psychometric functions is possible
- Intelligibility of noise vocoded speech
  - Relative criterion can vary by ~40 dB without loss of intelligibility
  - FFT length can vary between 20 and 60 ms without loss of int…
  - Uniform frequency resolution is better than non-uniform (erb)
  - Maximizing correlation in modulation domain is equivalent to HP filtering the spectrogram (when R=0)
    - Nice idea but little benefit
    - Seems logical to extend it to freq axis but gives small improvement

Can Binary Masks Improve Intelligibility?

- Replication of Kim et al (2009) show mask enhancement not straightforward to achieve
- Binary mask has two effects
  - Preserve speech information in noisy signal when SNR good enough
  - Encode speech information in vocoded noise when SNR poor
- Former just like any enhancement algorithm
- Latter relies on pattern recognition system
  - Which may perform badly at low SNR – just when it would be most useful