CCRMA/YAMAHA MASS Project
Masking Ambient Speech Sounds

Juan-Pablo Cáceres
MASS Project Background
Collaboration Summer 2006

**CCRMA**

Chris Chafe
Jonathan Berger
Jonathan Abel
Juan-Pablo Cáceres
Hiroko Terasawa
Jason Sadural

**Yamaha**
Center for Advanced Sound Technologies
Innovative Technology Group

Yasushi Shimizu
Atsuko Ito
The Problem

Unwanted sounds

Masking sound

speaker

Quiet room

Unwanted sounds

Mask/Camouflage of Intruding Speech
Some Maskers

Pink Noise Types, spectrally matched
   *Efficiency lost for speech spectrally unmatched*
   female+speech.wav

Amplitude Modulated Noise
   *Too distracting*
   AM_max45dB.wav

Reversed Speech
   *Too distracting*
   reverse+2dB.wav
Mass Project Summary

Tokyo Office Simulation for Experiment Setup

Masker Design: FM and Gurgle

Psychoacoustic Experiments
Recording for Room Modeling

- Sine Sweeps
- Voices
- Calibration Noise

Tokyo Office
Recording for Room Modeling

Tokyo Office
Impulse Response Generation

Recording

Calibration (EQ)

Room Modeling and Diffusion
Diffusion of Experiments
Speech Sounds in the Tokyo Office

Speech coming from the Adjacent Room

CONV_2people_e_adj_08dir1.wav

Speech Inside the Room

CONV_2people_e_ins_08dir1.wav

Need to be masked
Female Speech Analysis (Tokyo Office)

Roughly 3 main peaks bands

Spectrogram of Female 1

Spectrogram of Female 2

Spectrogram of Female 3
Male Speech Analysis (Tokyo Office)

Main Peak
Voice Comparison for One Word

Power Spectrum One Word Female Clip

Power Spectrum One Word Male Clip

Amplitude vs. Frequency (Hz)
Voice Analysis in the Tokyo Office

Wall: Low-Pass Filter

=> *Almost no energy above 1000 Hz*

Male voice: strong component around 100 Hz

Female main band: ~200/300 Hz
Stronger second band: ~400/600 Hz

*This is the Energy that we need to mask*
Masking Phenomenon

Loud sound & soft sound

Frequency Masking

Temporal Masking
Critical Bands

Frequency difference at which 2 pure tones can be easily distinguished ("roughness" disappears).

Stimulate the same section of the basilar membrane

Masking Efficiency Increase up to 1 CB
Perceptual Loudness inside 1 CB < less Perceptual Loudness spanning more than 1 CB

Noise is better masker than tones
Why Using Noise as Masker?

Is LESS distracting

Is MORE efficient

Use bands of noise of 1 Critical Bandwidth
**Noise Bands as Masker**

**ERB (Equivalent Rectangular Bandwidth) for each band**

\[ ERB = 24.7 \left( 4.37 \frac{f_c}{kHz} + 1 \right) \]
Cocktail Party Effect

Focus in *one* conversation in the midst of *other conversations* and *noise*

Simple and intuitive: **VASTLY** complex (physiological & technically)

Acoustic task:  
Separate out a single talker's speech from a complex spectrogram

Humans are extremely good at it
Factors that make the task easier

Directionality
Lip-Reading, gestures...
Voice quality (pitch, gender, ...)
Transition probabilities (i.e. MEANING)
Opposite problem

Jams a sound:

*Absorption into a cocktail party texture*

2 Approaches:

- Energetic+Information Masker: FM Noise
- Information Masker: “Gurgle”
Speech Like Modulation of Bands

Spectrogram, Random Walk Modulated Noise Example Prototype

Frequency [Hz]

Time (sec)

0 1 2 3 4 5 6 7 8 9
“Gurgle” Masker

Developed entirely by ear.

Highly-modulated information masker using *FM-speech synthesis*

`mask_gurgle.wav`
Experiment 1: Masker Refinement

General protocol: best masker from a parametric masker

Genetic algorithm approach
• Vary one parameter (all the others fixed)
• Find 1 or 2 “sweet spots”, fix that parameter
• Repeat the process for next parameter

Setup: Masker continuous stream, subject “reacts” when to intruding speech
Parameters for the 3 Noise Bands

- Center Frequency
- Relative Amplitude
- Modulation Rate
- Modulation Variance
Result examples after 1\textsuperscript{st} Stage (fc)
**FM Evolved Masker**

<table>
<thead>
<tr>
<th></th>
<th>Band 1</th>
<th>Band 2</th>
<th>Band 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>fc1 (dB)</td>
<td>106.8</td>
<td>210</td>
<td>420</td>
</tr>
<tr>
<td>Amp (dB)</td>
<td>-10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>fm (Hz)</td>
<td>2</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>fmAmpVar</td>
<td>1fc</td>
<td>1fc</td>
<td>3fc</td>
</tr>
</tbody>
</table>

*Note Harmonic Relationship in fc*
Experiment 2: Efficiency

Main Conclusions deal with directionality:

What if Intruding *Speech & Masker* come from the *same direction*?
Masking Noise Directionality

Keep source outside the room, remove wall effect.
Identify direct path, inverse filter it.

CONV_2people_e_adj_08dir1.wav

CONV_2people_e_adj_08dir1NW.wav
Gurgle Masker Directionality

Informal testing:

Directionality (spatialization)

Higher efficiency

speech+masker
Experiment 3: Annoyance

Effect on productivity tasks involving auditory/visual stimulation and response

Procedure:
• Word-list presented to the subject
• Masker is turned on (30 secs)
• Subject answers to mental math questions
• Subject try to recall the word-list (masker off)
# Some Results From Experiment 3

<table>
<thead>
<tr>
<th>Masker</th>
<th>Precision rate</th>
<th>Speed of task</th>
<th>Word recall rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>Baseline</td>
<td>Baseline</td>
<td>Baseline</td>
</tr>
<tr>
<td>FM</td>
<td>-14%</td>
<td>+10%</td>
<td>+18%</td>
</tr>
<tr>
<td>Gurgle</td>
<td>-0.5%</td>
<td>-17%</td>
<td>+43%</td>
</tr>
<tr>
<td>Mix</td>
<td>-22%</td>
<td>-28%</td>
<td>NA</td>
</tr>
<tr>
<td>MixALot</td>
<td>+8.5%</td>
<td>+2.4%</td>
<td>NA</td>
</tr>
</tbody>
</table>

*mask_control.wav  
*mask_fm.wav  
*mask_gurgle.wav  
*MaskMix.wav  
*Mask_Mix_a_lot.wav*
Conclusions and Future Work

1. We model real-world workplace
2. Intruding speech representative of real-world situations
3. Compare effectiveness & spatial arrangement
4. Compare the annoyance & degradation of mental tasks
Conclusions and Future Work

Two main type of maskers:

“Energy” Maskers & “Information” Maskers

Blend of both

Directionality (spatialization) seems to be of great importance
Questions? Discussion?