Progress in audible (normal) and inaudible speech recognition based on NAM microphones

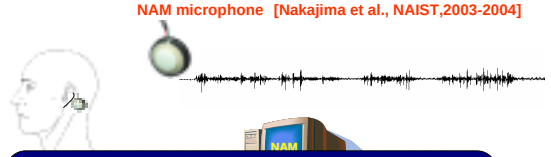
Panikos Heracleous

Speech and Acoustics Processing Lab.

Introduction

NAM microphone is a special device (sensor) which is attached behind the talker's ear and perceives very quietly uttered speech

NAM microphone [Nakajima et al., NAIST,2003-2004]



NAM plays an important role in Ubiquitous Networking
It extends the speech media applications to cover:
- **Applications with privacy (e.g. speech recognition, telephony)**
- **Applications under several (e.g. noisy) environments**
- **Applications for specific people groups (e.g. speech-impaired)**

NAM microphones

1. Stethoscopic NAM microphone [2003]
- Based on stethoscope used in medical science
- Able to capture audible and inaudible speech
- Only low frequency band
- Low speech intelligibility

Male

2. Silicon NAM microphone [2004]
- Silicon is used to wrap a sensitive microphone
- Able to capture audible and inaudible speech
- Wide frequency band
- Higher speech intelligibility

Male
Female

3. First pre-commercial prototype NAM [2004]
- Based on the Silicon NAM microphone
- Still under development and investigation

Some advantages of NAM

- Able to capture audible and inaudible speech
  - Privacy in communication (human machine, human-human)
  - Direct attachment to the skin – Expected noise robustness
- No additional techniques are required
  - Only a microphone pre-amplifier is required
  - Directly connected to PC for recording, or speech recognition
- Size, weight

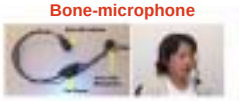
Music is heard

Silicon NAM microphone

A throat microphone

Some similar approaches

Bone-microphone



Microsoft

PARAT

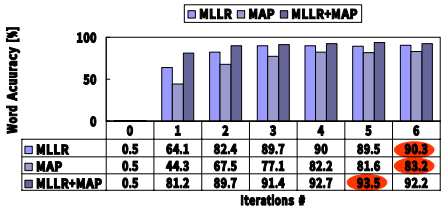
Norwegian University of Science and Technology

Throat microphones

- **Special applications**
  Military, Police, mobile telephone
- **Usually normal speech recognition**
  Integration with other sensors
- **Big size and uncomfortable usage**
- **Results show high noise robustness**

Stethoscopic NAM

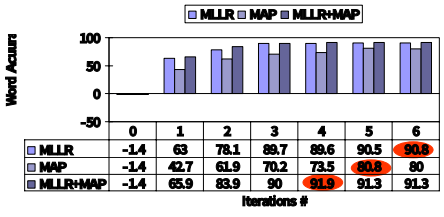
Speaker-dependent experiments (Clean)



	0	1	2	3	4	5	6
MLLR	0.5	64.1	82.4	89.7	90	89.5	90.3
MAP	0.5	44.3	67.5	77.1	82.2	81.6	83.2
MLLR+MAP	0.5	81.2	89.7	91.4	92.7	93.5	92.2

700 training utterances, 24 test utterances
Julius recognizer, 20k vocabulary Japanese dictation task
High performance! Comparable to normal-speech recognition

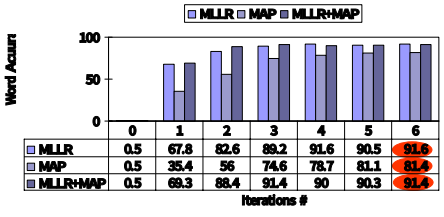
Stethoscopic NAM
Speaker-dependent experiments (Noise) - I



	0	1	2	3	4	5	6
MLLR	-1.4	63	78.1	89.7	89.6	90.5	90.8
MAP	-1.4	42.7	61.9	70.2	73.5	80.8	80
MLLR+MAP	-1.4	65.9	83.9	90	91.9	91.3	91.3

700 training utterances, 24 test utterances
Room noise is superimposed to clean data
Noise level: About 50dBA – HMMs noisy

Stethoscopic NAM
Speaker-dependent experiments (Noise) - II



	0	1	2	3	4	5	6
MLLR	0.5	67.8	82.6	89.2	91.6	90.5	91.6
MAP	0.5	35.4	56	74.6	78.7	81.1	81.4
MLLR+MAP	0.5	69.3	88.4	91.4	90	90.3	91.4

700 training utterances, 24 test utterances
Room noise is superimposed to clean data
Noise level: About 50dBA – HMMs clean

Stethoscopic NAM
Speaker-independent experiments

Word Accuracy [%]

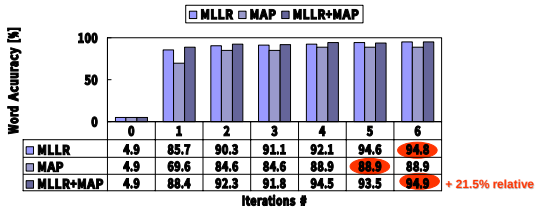
Speaker	Training method			
	MLLR	MAP	MLLR+ MAP	EM
Male	73.5	62.4	77.8	75.9
Female	70.0	58.4	71.1	65.4

Reasonable results

Recording conditions
- 16-bit SONY ZA5ES SUPER BIT MAPPING DAT tape digital recorder
- Anechoic room 30dBA, Sampling frequency: 44100kHz -> 16kHz

Recording data
-25 speakers (14 male, 11 female), 100 newspaper, 50 phoneme balanced utterances
-3189 training utterances, 48 training utterances

Silicon NAM
Speaker-dependent experiments (Clean)



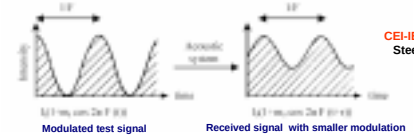
	0	1	2	3	4	5	6
MLLR	4.9	85.7	90.3	91.1	92.1	94.6	94.8
MAP	4.9	69.6	84.6	84.6	88.9	88.9	88.9
MLLR+MAP	4.9	88.4	92.3	91.8	94.5	93.5	94.9

+ 21.5% relative

350 training utterances, 24 test utterances
Julius recognizer, 20k vocabulary Japanese dictation task

Intelligibility objective measurement

The intelligibility of speech refers to the accuracy with which a normal listener can understand a spoken word or phrase



Modulated test signal — Received signal with smaller modulation

CEI-IEC standard n. 60268-216
Steeneken and Houtgast

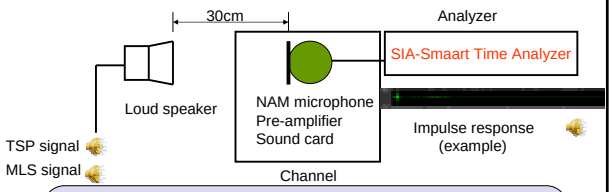
Speech Transmission Index (STI): Based on the reduction of m_i modulation index

Modulation function : Shows the reduction of the modulation index

$$m(F) = \frac{m_o}{m_i} \qquad m(F) = \frac{\int_0^\infty h^2{}_f(\tau).\exp(-j.2.\pi.F.\tau).d\tau}{\int_0^\infty h^2{}_f(\tau)d\tau}$$

Base on impulse response (Schroeder, 1981)

NAM microphone measurement



Speech Intelligibility objective measures

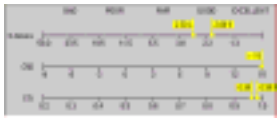
STI	Speech transmission index – Optimal 1.0 / Cannot hear 0.0
RSTI	Rapid Transmission Index – Data from narrow band
Clarity	Early to late energy ratio in impulse response
% AlCons	Articulation Loss of consonants

NAM microphones measurement

Investigating effects of design (e.g. wrapping), quality of NAM and recording setup correctness

TSP signal | Impulse response

1

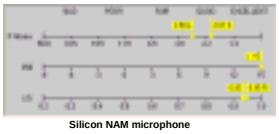


Reference microphone (Simple Desktop PC mic)

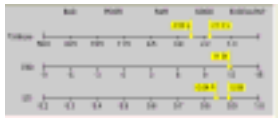
Pre-commercial Silicon NAM microphone

2

3



Silicon NAM microphone



Stethoscope microphone

4

Silicon microphones do not show significant degradations

Conclusions

- NAM microphones introduction
  - Able to capture audible and inaudible speech
  - Can be applied in speech recognition and communication
  - Easy to use – Small, light
- Experiments on audible speech recognition
  - Stethoscope NAM
    - Clean: 93.5% word accuracy
    - Noisy: 93.3% word accuracy (50dBA)
  - Silicon NAM
    - Clean: 94.86% word accuracy
- NAM quality and intelligibility investigation
  - NAM microphones show similar intelligibility to other sensors
- Data collection for speaker-independent experiments

Merry Christmas !