Improving Speaker Verification by Periodicity Based Voice Activity Detection

Ville Hautamäki (University of Joensuu, Finland),
Marko Tuononen (University of Joensuu, Finland),
Dr. Tuija Niemi-Laitinen (National Bureau of Investigation, Finland),
Prof. Pasi Fränti (University of Joensuu, Finland).

To be presented in the SPECOM'07.

AMICT'07
Petrozavodsk
21.08.2007

JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Prerequisities – Voice Activity Detector (VAD)

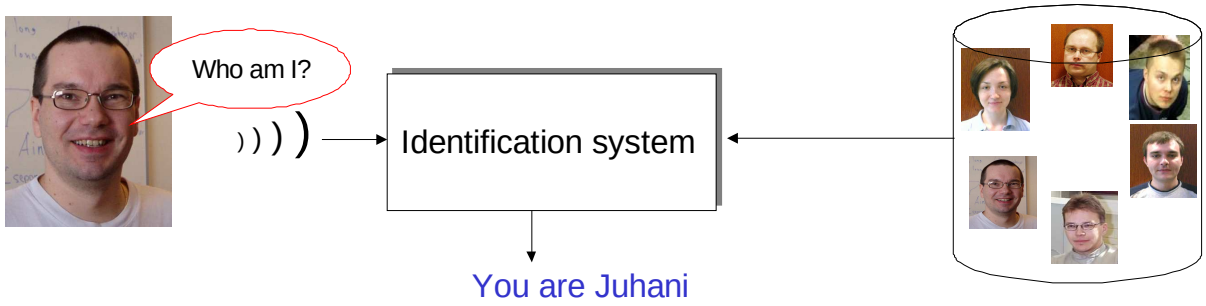
- Basic idea: Classifies a given sound frame as a speech or as a non-speech.
- Needed in most speech technology applications. For example in
  - speech recognition/enhancement, and
  - voice biometric.
- No “one-solution-fits-all” exists.

JOENSUUN YLIOPISTO

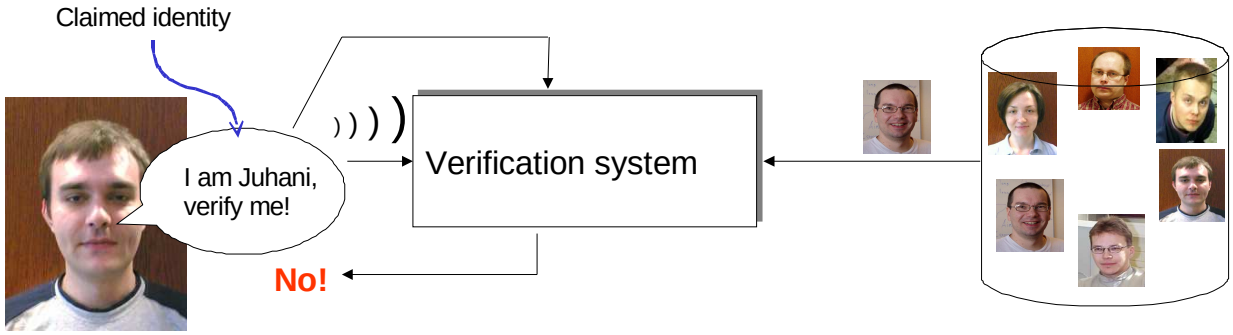
Joensuun yliopisto
PL 111
80101 Joensuu
puh. (013) 251 111
fax (013) 251 2050
www.joensuu.fi

Prerequisities – Speaker recognition

Identification: “Whom this voice belongs to?”



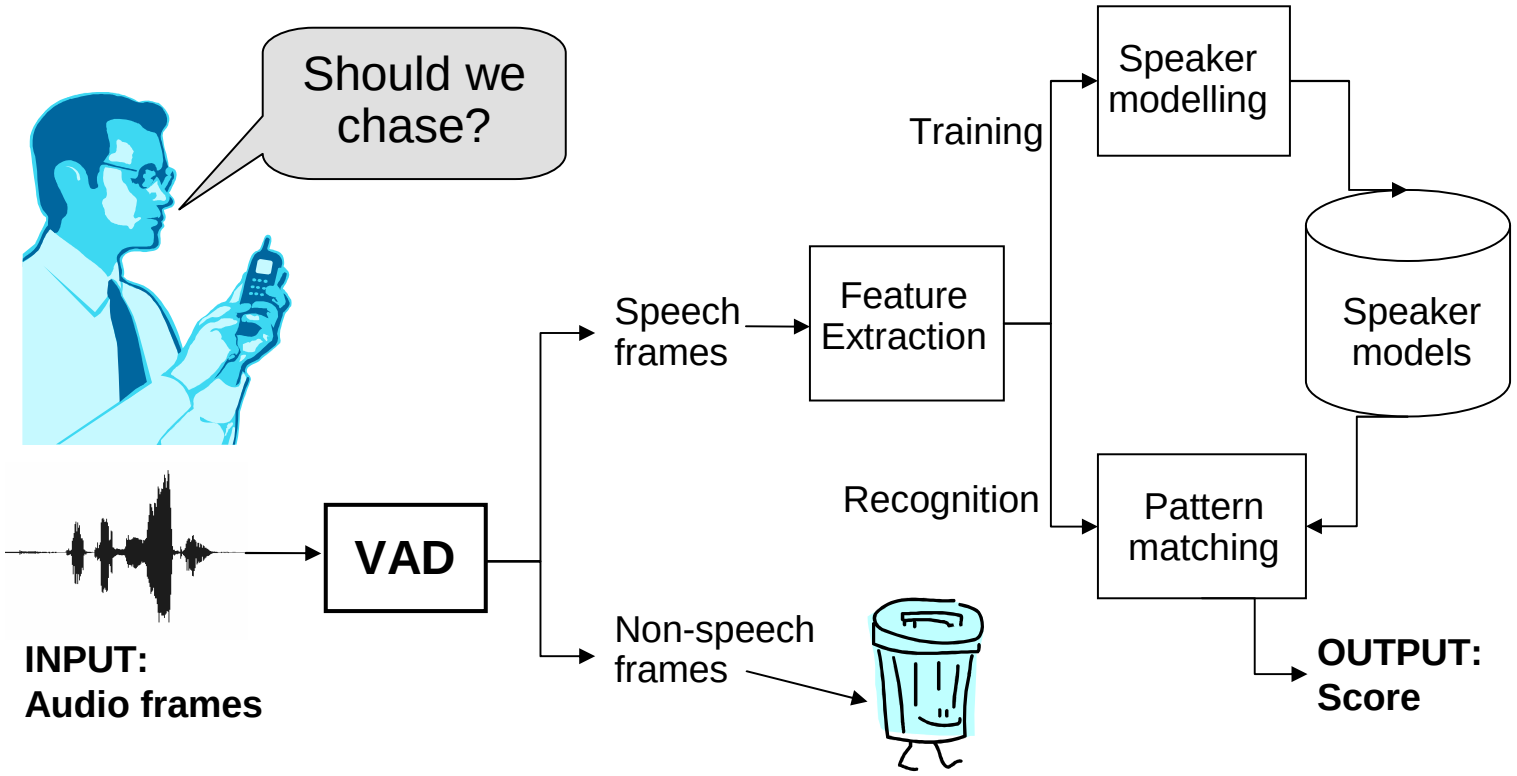
Verification: “Is this Juhani's voice?”



Joensuun yliopisto
PL 111
80101 Joensuu
puh. (013) 251 111
fax (013) 251 2050
www.joensuu.fi

Prerequisities – VAD in Speaker verification system

- VAD is used as as a preprocessor for the **realtime** speaker verification.



JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Motivation

- Why VAD? Including non-speech frames will bias the speaker model.
- Why periodicity-based VAD? Voiced phonemes more discriminative than unvoiced ones (Sambur, 1975).
- Online or offline? In realtime speaker verification, we need to start processing speech frames with as small delay as possible.

JOENSUUN YLIOPISTO

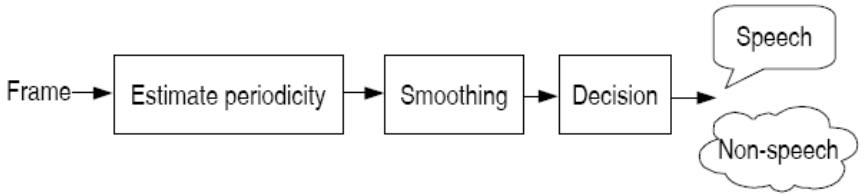
Joensuun yliopisto
PL 111
80101 Joensuu
puh. (013) 251 111
fax (013) 251 2050
www.joensuu.fi

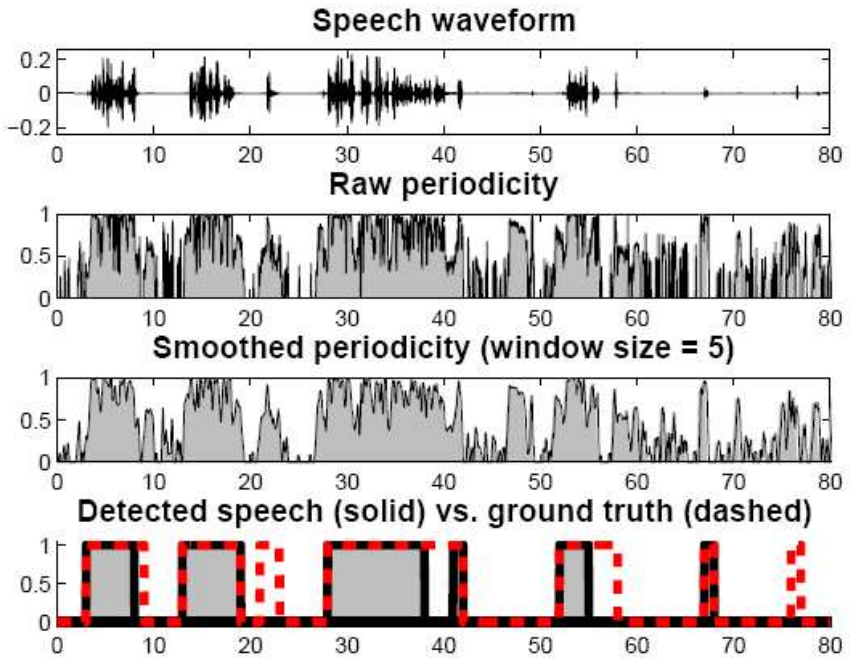
Periodicity VAD (proposed)

- Assumption: speech is periodic.
- Periodicity is extracted from each frame using YIN pitch estimation algorithm (Cheveigne and Kawahara, 2002).

$$d(\tau) = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2$$

- Realtime!





Energy-based VAD

- Simple energy-based approach that was used in NIST-06 eval (Tong et al., 2006).
- Measures intra-frame energy by calculating standard deviation of the frame.
- Not realtime.

$$S_i = 20 \log_{10} \sqrt{\frac{1}{N-1} \sum_{j=1}^{W} (x_j^{(i)} - \bar{x}^{(i)})^2},$$

Then frame i is detected as speech
if $S_i > (\max_j S_j - T)$ and $S_i > -55$.

JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Long Term Spectral Divergence VAD

- Compares the long-term spectral envelope to the average noise spectrum (Ramirèz et al., 2004).
- Realtime!

$$\mathrm{LTSE}_M(k,l) = \max_{j=-M}^{M} X(k,l+j),$$

$$\mathrm{LTSD}_M(l) = 10\log_{10}\left(\frac{1}{NS}\sum_{k=0}^{NS-1}\frac{\mathrm{LTSE}^2(k,l)}{N^2(k)}\right)$$

JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Speaker verification experiments
test setup

- Frames: 30ms with 33% overlap.
- Features: MFCC coefficients 1-12 calculated from 27 channel mel-filterbank, appended with its delta and double-delta coefficients, 0/1-normalized.
- Models: adapted Gaussian mixture models (Reynolds et al., 2000).

JOENSUUN YLIOPISTO

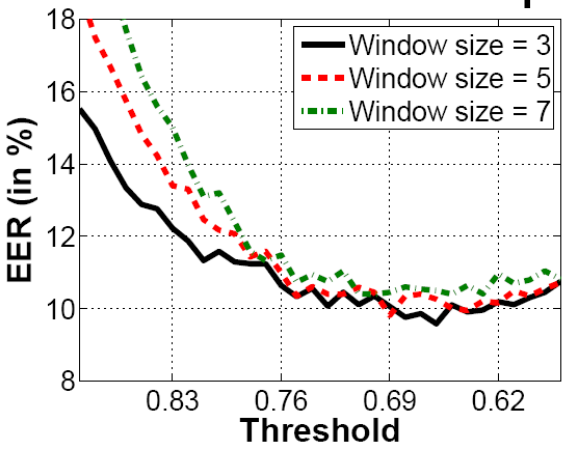
Joensuun yliopisto
PL 111
80101 Joensuu

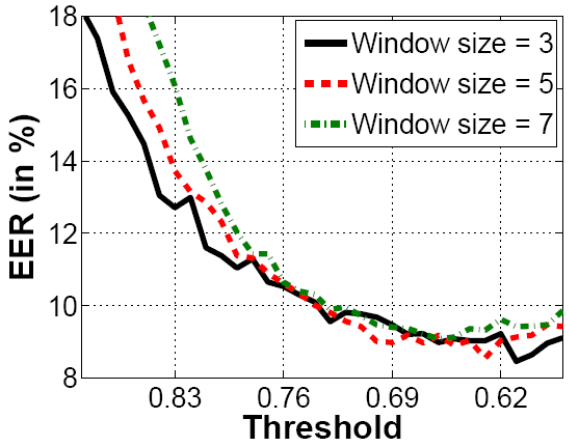
puh. (013) 251 111
fax (013) 251 2050

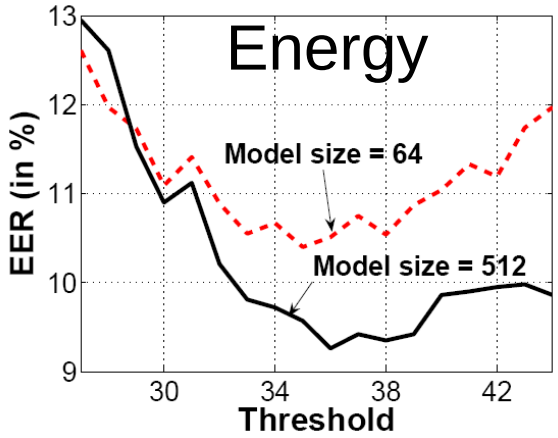
www.joensuu.fi

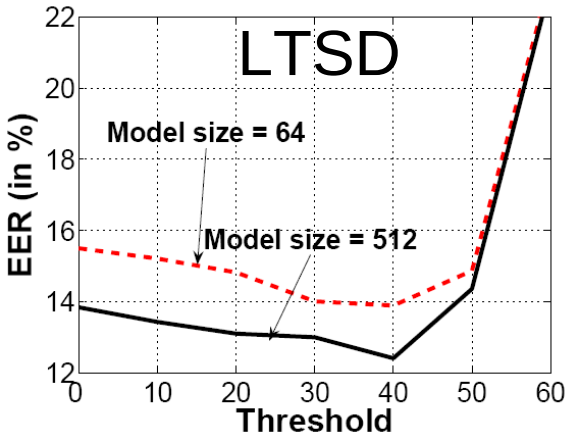
Speaker verification experiments
tuning on NIST-01 evalset

Proposed method









Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Speaker verification experiments
results on NIST-06 1conv-1conv

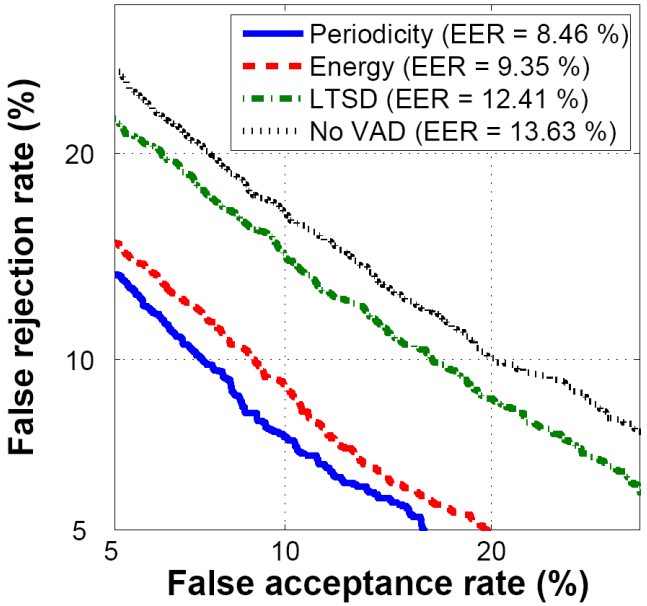


Figure 5: NIST 2001 best results for the model size 512.

Table 1: Summary of speaker verification results (% EER).

	NIST 2001				NIST 2006
	model 512		model 64		model 512
	EER	Thr	EER	Thr	EER
No VAD	13.63		16.00		44.39
LTSD	12.41	40	13.74	45	35.82
Energy	9.26	36	10.40	35	**16.63**
Periodicity	**8.46**	0.61	**9.58**	0.66	16.76

JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Speech segmentation experiments

- Segment the audio signal in time to alternating speech and non-speech blocks.
- For evaluating segmentations, we have formed ground correct segmentation manually using one second resolution.

	Bus-stop	Lab	NIST05
LTSD adaptive	19	14	40
LTSD trained	**6**	15	**1**
Energy	15	17	2
Periodicity	21	**10**	3

JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

Summary

- Proposed VAD is based on realtime periodicity analysis
  - part of the speech is periodic.
- Outperforms LTSD method, comparably good with energy-based method.
- Offers improvement to the realtime speaker verification system.

JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi

EOF

- Questions?

JOENSUUN YLIOPISTO

Joensuun yliopisto
PL 111
80101 Joensuu

puh. (013) 251 111
fax (013) 251 2050

www.joensuu.fi