INTERSPEECH 2010
MAKUHARI JAPAN 26-30 SEPTEMBER
Spoken Language Processing for All
International Speech Communication Association

Thu-Ses2-P4

-

Voice activity and turn detection

Listen! 30th Nov 2010

Erich Zwyssig

Introduction

- VAD/SAD (Speech/Voice Activity Detection)
  - Chair: Tomohiro Nakatani NTT (Communication Science Laboratories)
  - 14 posters (look at 2, plus one reference)
    - Sohn et al. (Goal Post)
    - Ghaemmaghami et al.
    - QUT-Noise-Timit Corpus

A Statistical Model-Based Voice Activity Detection

Jongseo Sohn, *Student Member, IEEE,* Nam Soo Kim, *Member, IEEE,* and Wonyong Sung

***Abstract*— In this letter, we develop a robust voice activity detector (VAD) for the application to variable-rate speech coding. The developed VAD employs the decision-directed parameter estimation method for the likelihood ratio test. In addition, we propose an effective hang-over scheme which considers the previous observations by a first-order Markov process modeling of speech occurrences. According to our simulation results, the proposed VAD shows significantly better performances than the G.729B VAD in low signal-to-noise ratio (SNR) and vehicular noise environments.**

***Index Terms*— Decision-directed estimation, hidden Markov model, likelihood ratio test, voice activity detection.**

Likelihood Ratio Test

H_0: speech absent: $\mathbf{X} = \mathbf{N}$

H_1: speech present: $\mathbf{X} = \mathbf{N} + \mathbf{S}$

$$p(\mathbf{X}|H_0) = \prod_{k=0}^{L-1} \frac{1}{\pi\lambda_N(k)} \exp\left\{-\frac{|X_k|^2}{\lambda_N(k)}\right\}$$

$$p(\mathbf{X}|H_1) = \prod_{k=0}^{L-1} \frac{1}{\pi[\lambda_N(k) + \lambda_S(k)]} \cdot \exp\left\{-\frac{|X_k|^2}{\lambda_N(k) + \lambda_S(k)}\right\}$$

$$\log \Lambda = \frac{1}{L} \sum_{k=0}^{L-1} \log \Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta.$$

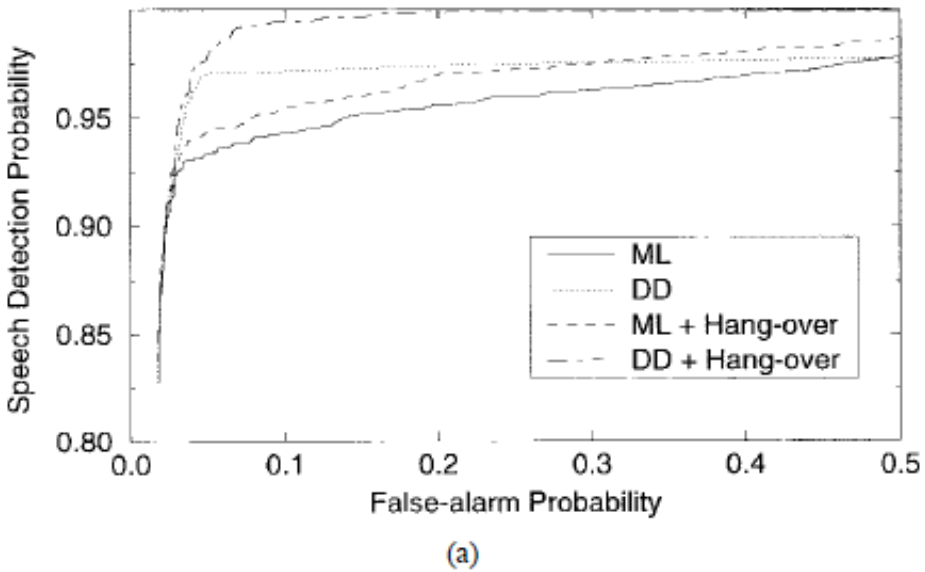
HMM-based Hang-Over Scheme

In this Markovian frame state model, the current state depends on the previous observations as well as the current one, which is reflected on the decision rule in the following way:

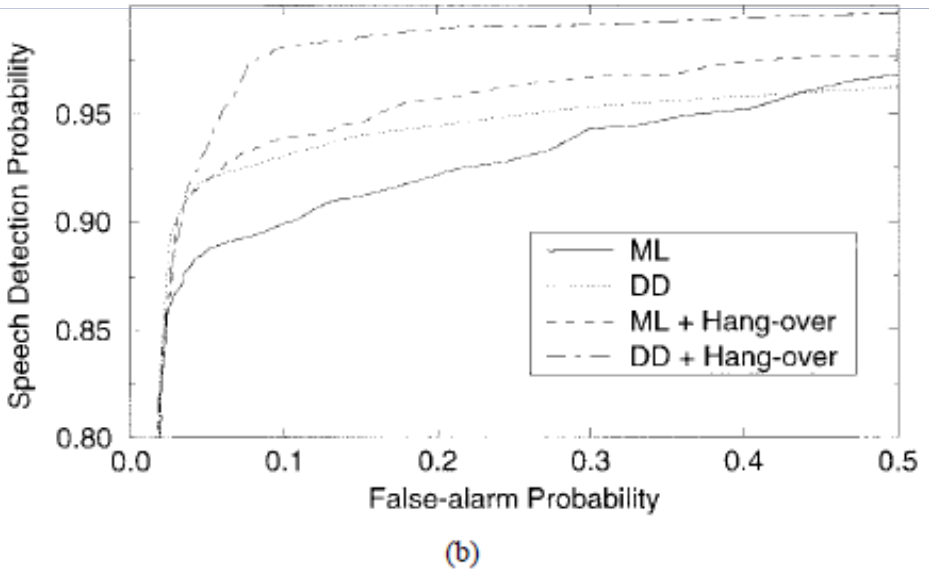
$$\mathcal{L}(n) \triangleq \frac{p(\mathcal{X}_n|q_n = H_1)}{p(\mathcal{X}_n|q_n = H_0)} = \frac{P(H_0)}{P(H_1)} \frac{P(q_n = H_1|\mathcal{X}_n)}{P(q_n = H_0|\mathcal{X}_n)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \qquad (9)$$

where $\mathcal{X}_n = \{\mathbf{X}(n), \mathbf{X}(n-1), \cdots, \mathbf{X}(1)\}$ represents the set of observations up to the current frame n. For the effi-

Results



(a)



(b)

Fig. 1. Receiver operating characteristics of ML and DD based decision rules with and without hang-over at 5 dB SNR. (a) Vehicular noise. (b) White noise.

Noise Robust Voice Activity Detection Using Features Extracted From the Time-Domain Autocorrelation Function

Houman Ghaemmaghami, Brendan Baker, Robbie Vogt, Sridha Sridharan

Speech and Audio Research Laboratory, Queensland University of Technology, Brisbane, Australia

Abstract

This paper presents a method of voice activity detection (VAD) for high noise scenarios, using a noise robust *voiced* speech detection feature. The developed method is based on the fusion of two systems. The first system utilises the maximum peak of the normalised time-domain autocorrelation function (MaxPeak). The second system uses a novel combination of cross-correlation and zero-crossing rate of the normalised autocorrelation to approximate a measure of signal pitch and periodicity (CrossCorr) that is hypothesised to be noise robust. The score outputs by the two systems are then merged using weighted sum fusion to create the proposed autocorrelation zero-crossing rate (AZR) VAD. Accuracy of AZR was compared to state-of-the-art and standardised VAD methods and was shown to outperform the best performing system with an average relative improvement of 24.8% in half-total error rate (HTER) on the QUT-NOISE-TIMIT database created using real recordings from high-noise environments.

Index Terms: voice activity detection, high noise, autocorrelation, zero-crossing rate, time-domain analysis

MaxPeak Algorithm

After pre-emphasis, the normalised time-domain autocorrelation, $R_k[z]$, at lags corresponding to pitch periods of 2 to 20 ms, is calculated for $x_k[i]$,

$$R_k[z] = \frac{\sum_{i=1}^{n-z} x_k[i]x_k[i+z]}{\sum_{i=1}^{n} x_k^2[i]} \quad (3)$$

where z is the autocorrelation lag and n is the number of samples in $x_k[i]$, hence, the MaxPeak score ($M[k]$), for the k^{th} frame, $x_k[i]$, is calculated as,

$$M[k] = \max(R_k[z]) \quad (4)$$

it is expected that voiced speech would produce a higher maximum peak value than unvoiced or silence/noise frames.

CrossCorr Algorithm

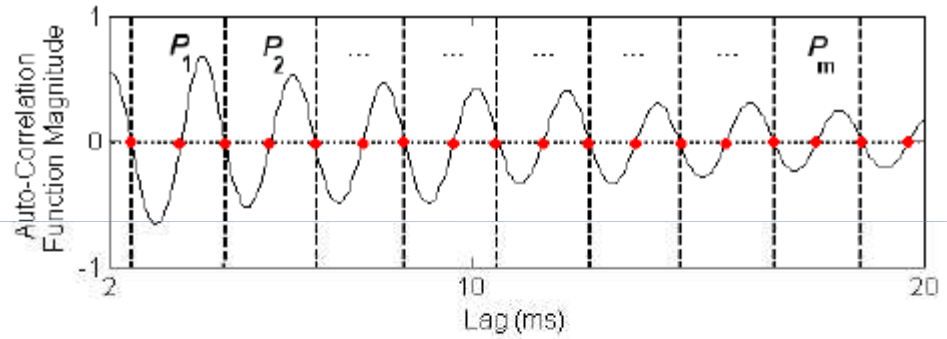


Figure 1: $R_k[z]$ *of a voiced speech frame for 2 to 20 ms lag. The CrossCorr algorithm initially performs a count of the marked zero-crossing points to estimate the pitch and perform a cross-correlation similarity check of adjacent* P_y *segments for* $y = 1, 2, \ldots, m$.

Figure 1 shows that once the zero-crossing rate of $R_k[z]$, which indicates the approximate pitch of the analysed frame, satisfies the specified upper and lower limits (it is within the approximate 50 to 500 Hz pitch period) the CrossCorr, $C[k]$, feature can then be calculated using (5) and (6),

$$\hat{R}_y[z'] = \sum_{j=1}^{n'-z'} P_y[j] P_{y+1}[j + z'] \quad (5)$$

therefore,

$$C[k] = \sum_{y=1}^{m-1} \max(\hat{R}_y[z']) \quad (6)$$

where P_y specifies an assumed "period" of $R_k[z]$, and $\hat{R}_y[z']$ is the cross-correlation function between P_y and its posterior "period". This adjacent cross-correlation is used to ensure maximum CrossCorr score value in cases such as that in Figure 1, where the magnitude of the autocorrelation decreases with lag increase.

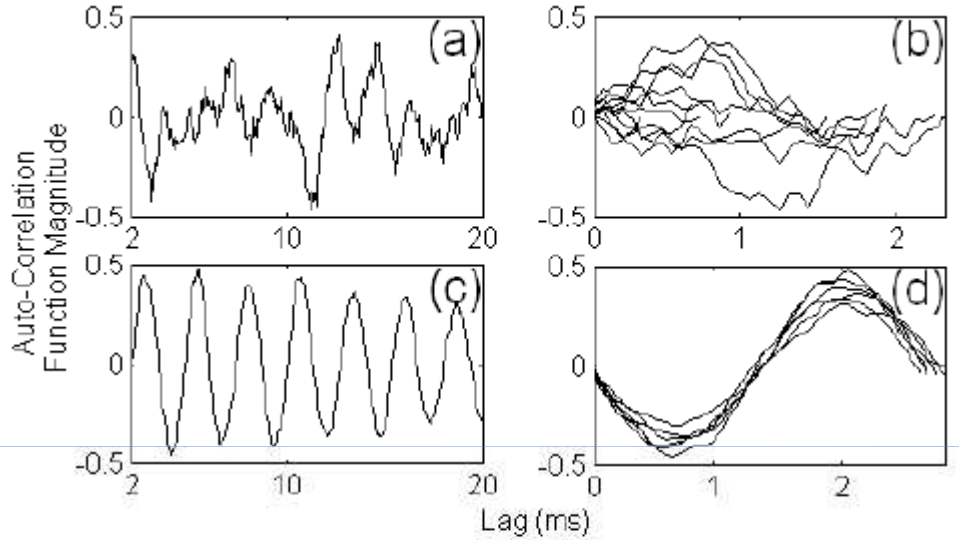


Figure 2: *(a) $R_k[z]$ of a noise frame (2 to 20 ms lag) from a noisy speech file at SNR=(-5)dB. (b) Similarity check of P_y "periods" every second zero-crossing indicates low correlation. (c) $R_k[z]$ of a voiced speech frame at SNR=(-5)dB. (d) Similarity check of P_y "periods" indicates high correlation.*

Results

Table 2. *Overall %FAR, %MR, and %HTER for the AZR VAD and baseline systems at each tested noise level.*

VAD Systems	Low Noise (SNR=10 or 15dB)			Medium Noise (SNR=0 or 5dB)			High Noise (SNR=-10 or -5dB)		
	% FAR	% MR	% HTER	% FAR	% MR	% HTER	% FAR	% MR	% HTER
AZR	15.6	6.6	**11.1**	20.5	12.1	**16.3**	31.9	25.5	**28.7**
LTSD	20.7	12.8	**16.7**	26.2	19.5	**22.8**	28.6	36.8	**32.7**
Sohn's (LRT)	24.6	20.4	**22.5**	33.5	28.9	**31.2**	56.5	25.0	**40.8**
G.729-B	33.7	18.8	**26.2**	34.2	31.5	**32.9**	35.0	50.2	**42.6**
ETSI	68.3	0.2	**34.2**	66.8	2.2	**34.5**	65.1	13.6	**39.4**

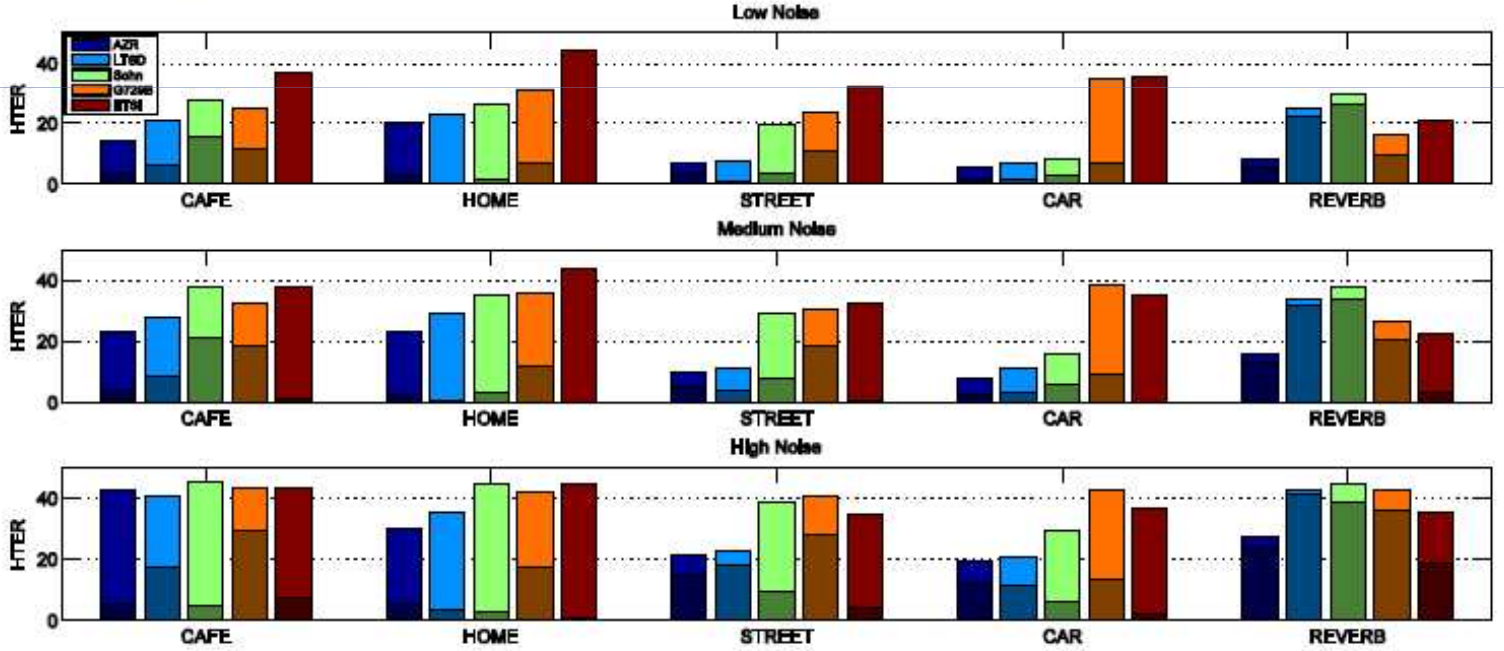


Figure 3: *%HTER performance of AZR VAD and baseline methods for each noise scenario at three tested noise levels. The dark shading of each bar represents the %MR and lighter shade displays the %FAR of the overall %HTER.*

The QUT-NOISE-TIMIT Corpus for the Evaluation of Voice Activity Detection Algorithms

David Dean, Sridha Sridharan, Robert Vogt, Michael Mason

Speech and Audio Laboratory, Queensland University of Technology
Brisbane, QLD, Australia

Abstract

The QUT-NOISE-TIMIT corpus consists of 600 hours of noisy speech sequences designed to enable a thorough evaluation of voice activity detection (VAD) algorithms across a wide variety of common background noise scenarios. In order to construct the final mixed-speech database, a collection of over 10 hours of background noise was conducted across 10 unique locations covering 5 common noise scenarios, to create the QUT-NOISE corpus. This background noise corpus was then mixed with speech events chosen from the TIMIT clean speech corpus over a wide variety of noise lengths, signal-to-noise ratios (SNRs) and active speech proportions to form the mixed-speech QUT-NOISE-TIMIT corpus. The evaluation of five baseline VAD systems on the QUT-NOISE-TIMIT corpus is conducted to validate the corpus and show that the variety of noise available will allow for better evaluation of VAD systems than existing approaches in the literature.

Index Terms: voice activity detection, speech databases, evaluation protocols

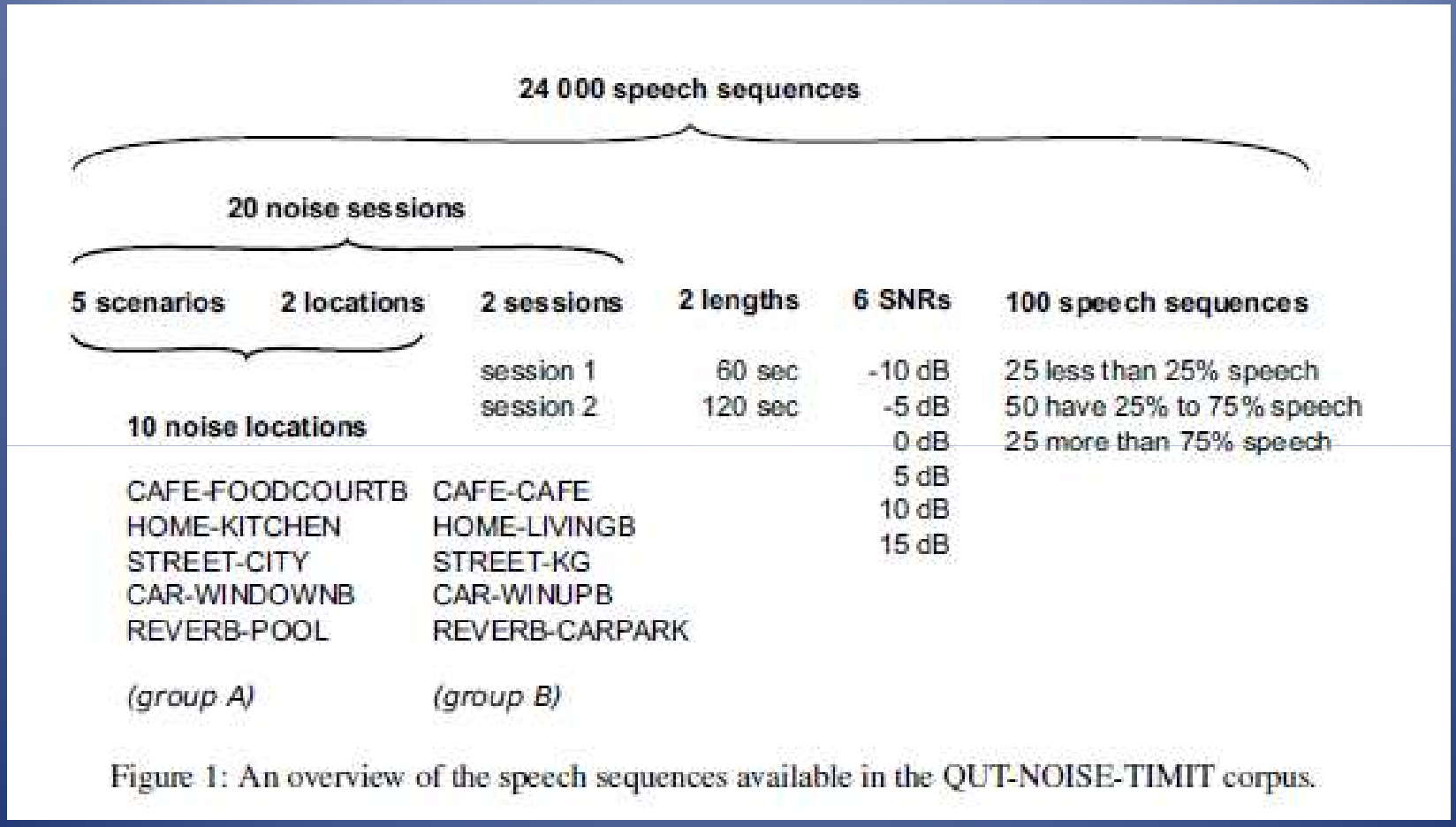


Figure 1: An overview of the speech sequences available in the QUT-NOISE-TIMIT corpus.

Thanks

- Any questions?

Turn Taking-based Conversation Detection by Using DOA Estimation

Yohei Kawaguchi, Masahito Togami, and Yasunari Obuchi

Central Research Laboratory, Hitachi, Ltd.
1-280, Higashi-koigakubo Kokubunji-shi, Tokyo 185-8601, Japan

Abstract

We propose a new method that detects conversation groups when multi-conversation groups exist simultaneously. The proposed method uses hands-free microphone arrays without wearable microphones. It has two main features: (a) We integrate a conventional turn taking-based conversation detection method with Direction of Arrival (DOA) estimation-based Voice Activity Detection (VAD). (b) The proposed method estimates the number of speakers for DOA estimation-based VAD by using the turn-taking rules. Experimental results indicate that the performance of the proposed method with only microphone arrays setup in rooms is comparable to that of the conventional methods with wearable microphones.

Index Terms: microphone array, direction of arrival, conversation detection, turn taking, voice activity detection.