Speech Event Detection using Multiband Modulation Energy

Georgios Evangelopoulos
Petros Maragos

Computer Vision, Speech Communication
& Signal Processing Group,
School of Electrical and Computer Engineering,
National Technical University of Athens,
Greece

URL: *http://cvsp.cs.ntua.gr*

1. Summary-Motivations

- Speech detection applications:
  - Recognition (pattern formation, frame dropping)
  - Processing (reduction, labeling)
  - Transmission, Coding
  - Enhancement, denoising, normalization
- Non-linear speech modeling, speech modulations
- Energy and spectral features via multiband analysis and dominant modulation energy tracking.
- Speech events in noise:
  endpoint detection (40% error reduction on NTIMIT)
  voice activity detection (7.5% on Aurora 2 and 9.5% on Aurora 3 error reduction)

2. Energy Operators & Modulations

- AM-FM modulation speech model.
- Real-valued AM-FM signal $x(t) = a(t)\cos\left(\int_0^t \omega(\tau)d\tau\right)$
  $a(t)$, $\omega(t)$: time-varying amplitude envelope and instantaneous frequency signals.
- Non-linear Teager -Kaiser differential energy operator for continuous-time signals:

$$\Psi[x(t)] \equiv [\dot{x}(t)]^2 - x(t)\ddot{x}(t)$$

  where $\dot{x}(t) = dx(t)/dt$.
- $\Psi[x(t)] \approx a^2(t)\omega^2(t) \rightarrow$ instantaneous source energy.
- *Energy Separation Algorithm* (ESA):

$$\sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}} \approx \omega(t) \quad , \quad \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \approx |a(t)|$$

3. Detection-theoretic Analysis

- Speech model $s[n] = \sum_{k=1}^{K} A_k[n]\cos(\Omega_{ck}\cdot n + \Phi_k[n])$, with k the resonance index of K speech formants.
- Multiple Hypothesis Testing to detect one out of many sinusoids of frequency Ω_{ck} or white gaussian background noise.
- $H_0 : X[n] = W[n],\ H_k : X[n] = W[n] + A_k\cos(\Omega_{ck}\cdot n + \Phi_k) + B_k$, for each frame of length N and $k = 1\cdots K$.
- Rule for speech detection ($\sigma_g \sim$ filter bandwidth):

$$N\frac{\hat{A}_k^2}{4\hat{\sigma}_1^2} + \ln\left(\Omega_{ck}^2 + \sigma_g^2\right) \underset{H_0}{\overset{H_k}{\gtrless}} \mathcal{O}(\hat{B}_k, \hat{\sigma}_k^2, \hat{\sigma}_0^2, N)$$

- Expected log value of Ψ on a filtered AM-FM plus noise

$$\ln\Psi(X[n]) \approx \ln A_k^2 + \ln(\Omega_{ck}^2 + \Gamma_{ck}) + \text{const.}$$

  Γ_{ck} constant of averaged filtered noise power.

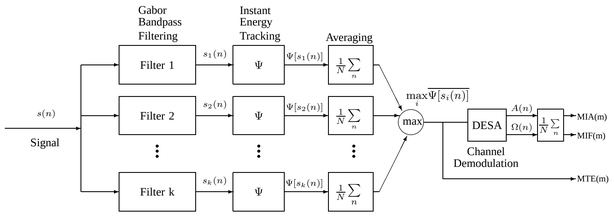
4. Multiband Modulation Features I

- Modulation bands are obtained through a linearly-spaced bank of K Gabor bandpass filters.
- Nonlinear energy measurement via the discrete Ψ_d.
- Tracking the *maximum average Teager Energy* (MTE):

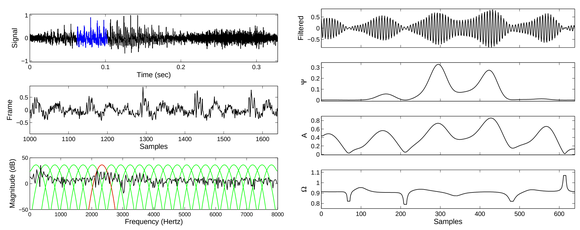
$$\text{MTE}(m) = \max_{1\le k\le K}\frac{1}{N}\sum_{n=1}^{N}\Psi_d[(s * h_k)(n)]$$

  h_k is the impulse response of k-th filter.
- ESA demodulation of filter $j = \arg\max_k(\text{MTE})$.
- *mean Multiband Instant Amplitude* (MIA) and *Instant Frequency* (MIF) features.
- MTE $\rightarrow$ *dominant* signal *modulation energy*.

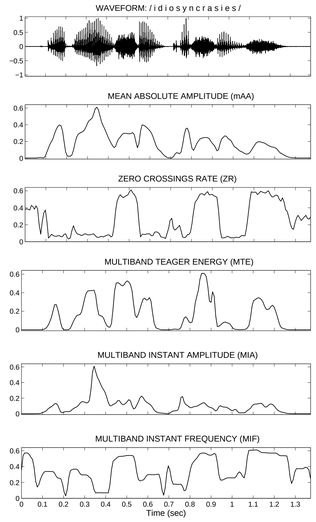
4. Multiband Modulation Features II



Max Energy Band demodulation



5. Speech Analysis Features



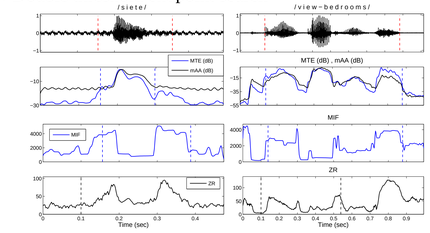
6. Speech Endpoint Detection

- Modulation features instead of conventional *mean absolute amplitude* (mAA) and *zero-crossings rate* (ZR).
- Noise statistics from first 100ms (mean μ_{sif}, standard deviation σ_{sif} of MIF and max MTE values $S_{\max}$ for 'silence' and for the whole signal $P_{\max}$).
- Threshold rules:

$$\gamma_{\text{f}} = \mu_{\text{sif}} + \kappa\sigma_{\text{sif}},\ \gamma_{\text{d}} = \min(T_1, T_2),\ \gamma_{\text{u}} = 5\cdot\gamma_{\text{d}}$$

$$T_1 = \lambda P_{\max} + (1-\lambda)S_{\max} \quad , \quad T_2 = 3\cdot S_{\max}$$

  κ, λ are weighting constants.
- Double threshold endpoint search



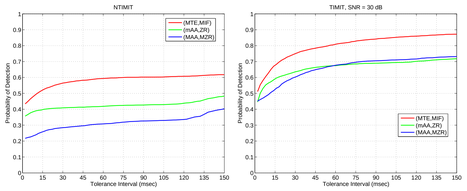
7. Experimental Evaluation

Percentage (%) of detected speech endpoints for various feature combinations on NTIMIT (1680 utterances)

Classic		Multiband	Multiband Modulation		Teager	
mAA, ZR	mSA, ZR	MAA, MZR	**MTE, MIF**	MTE	STE	PTE
56.1	66.6	51.5	73.5	73.1	71.6	49.5

Error: Boundary misplacement > 60 ms.

Detection-Error Tolerance Curves



Lost phoneme tolerance: 30 ms.

8. Voice Activity Detection

- *Long-Term Spectral Divergence* (LTSD) feature for VAD [Ramirez, Segura et. al. 2004].

  Level difference from background noise is compared to an adaptive threshold:

$$\gamma = \gamma_0 + (\gamma_1 - \gamma_0)(E - E_0)/(E_1 - E_0)$$

  E: background noise energy, E_0 in clean and E_1 in noisiest conditions.

- Modulation Energy Detection:
  1) *Multiband Teager Energy Divergence*:

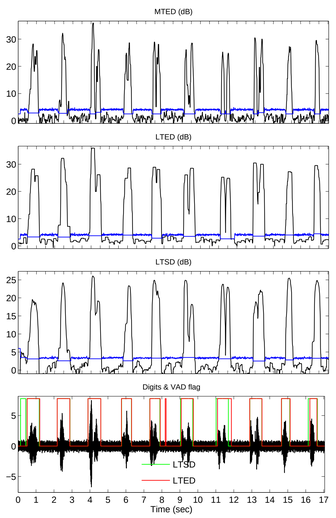
$$\text{MTED}(m) = 10\log_{10}(\text{MTE}(m)/\text{MTEW})$$

  MTEW the feature for background noise.
  2) *Long-term Multiband Teager Energy Divergence*:

$$\text{LTED}(m) = 10\log_{10}\left(\max_l\{\text{MTE}(m+l)\}/\text{MTEW}\right)$$

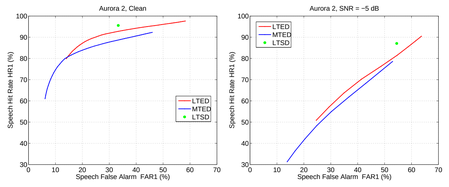
  $-L \le l \le L$ defines the order of dependence.

9. VAD Modulation Features



10. Experimental Framework

- Detection performance of the MTE-based VADs, on Aurora 2 (70070 utterances) and Aurora 3 (4914 utterances) in terms of errors at various SNRs.
- ASR to label speech and silence events. High recognition rate results (99.6% and 93.7% respectively) on the clean set define ground truth.
- Optimum decision thresholds by ROC curves (minimum distance from upper left corner).

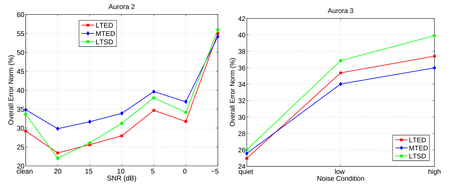


- Aurora 2 thresholds in dB
  MTED: $\gamma_0 = 24, \gamma_1 = 0.5$, LTED: $\gamma_0 = 32, \gamma_1 = 2$
- Aurora 3 $\gamma_0 = 6, \gamma_1 = 2.5$ for all three features.

11. VAD Evaluation

- Hit Rates: Speech $\text{HR1} = \frac{\text{Ds}}{\text{Ts}}$, Non-speech $\text{HR0} = \frac{\text{Dn}}{\text{Tn}}$
  D: # detected frames, T: # true frames, s: speech, n: noise
- Minimize the *overall false alarm error norm*:

$$\text{E}_{\text{FAR}} = \|(\text{FAR0,FAR1})\| = \left[(1-\text{HR0})^2 + (1-\text{HR1})^2\right]^{1/2}$$



- Aurora 2:
  Average HR $> 70\%$ for MTE-based algorithms.
  LTED i) minimizes E_{FAR}, ii) decreases 7.6% LTSD.
- Aurora 3:
  LTED i) higher HR than LTSD, ii) decreases 7.7% LTSD.
  MTED i) minimizes E_{FAR}, ii) decreases 9.5% LTSD.

References

M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 2, pp. 109–118, Feb. 2002.

J. Ramirez, J. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3-4, pp. 271–287, Apr. 2004.

G. Ying, C. Mitchell, and L. Jamieson, "Endpoint detection of isolated utterances based on a modified Teager energy measurement," in *Proc. IEEE ICASSP '93*, Minneapolis, MN, Apr. 1993, pp. 732–735.

J. Ramirez, J. Segura, C. Benitez, A. de la Torre, and A. Rubio, "A new Kullback-Leibler VAD for speech recognition in noise," *IEEE Signal Processing Lett.*, vol. 11, no. 2, pp. 266–269, Feb. 2004.

A. Bovik, P. Maragos, and T. Quatieri, "AM-FM energy detection and separation in noise using multiband energy operators," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3245–3265, Dec. 1993.

P. Maragos, J. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Processing*, vol. 41, no. 10, pp. 3024–3051, Oct. 1993.

L. Rabiner and M. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Systems Tech. J.*, 54: 2, pp. 297–315, Feb. 1975.

F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano, "Performance evaluation and comparison of G.729/AMR/Fuzzy voice activity detectors," *IEEE Signal Processing Lett.*, vol. 9, no. 3, pp. 85–88, Mar. 2002.

G. Evangelopoulos and P. Maragos, "Multiband Modulation Energy Tracking for Noisy Speech Detection ," *IEEE Trans. Speech Audio Processing, to appear*, 2005.