DySANA: Dynamic Speech and Noise Adaptation for Voice Activity Detection

Ron J. Weiss[1], **Trausti Kristjansson**[2]

[1]LabROSA · Dept of Electrical Engineering · Columbia University, New York, USA
[2]Google Inc., New York, USA
ronw@ee.columbia.edu, trausti@google.com

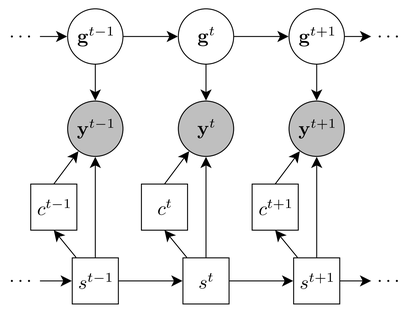
1. Summary

- Task: Speech endpoint detection in the presence of non-stationary noise.
- Combine adaptive energy threshold with classification approach by adapting classification model to signal-to-noise ratio conditions of observed signal.
- Simultaneously track signal and noise levels using a dynamic Bayesian network.
- Speech decision is made by combining SNR-dependent features with spectral shape features invariant to signal level.

2. Noisy Speech Signal Model

- Assume each frame is dominated by either speech or noise.
- Model both speech and noise using Gaussian mixture models with diagonal covariances.
- Use MFCC features so most feature dimensions are independent of signal energy.
- Track speech and noise gains relative to the corresponding models to adapt energy-dependent portions of the model (i.e. C_0 features) to environmental conditions.

Figure: DySANA graphical model



- Kalman filter with switched observations
- Hidden Markov model smoothing of speech/nonspeech decision
- Speech likelihood:

$$P(\mathbf{y}^t|s^t=1,\mathbf{y}^{1..t-1}) = \sum_c \pi_c \mathcal{N}\left(\mathbf{y}^t;\ \mu_{xc}+\ell\mu_{g_x^t},\ \Sigma_{xc}+\ell\ell^T\sigma_{g_x^t}\right)$$

where $\ell = [1\ \mathbf{0}]^T$

3. Model Dynamics

Compare different distributions for propagating gain distribution across time.

1. Random walk dynamics

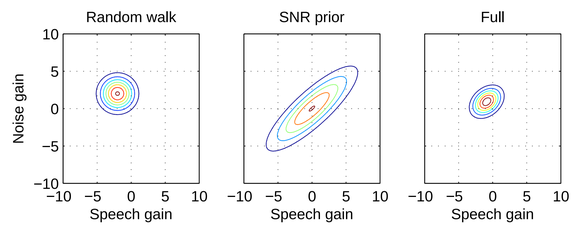
$$P(\mathbf{g}^{t+1}|\mathbf{g}^t) = \mathcal{N}(\mathbf{g}^{t+1};\ \mathbf{g}^t,\ \Sigma_{RW})$$

- State variables can move in any direction from current estimate. Σ_{RW} controls rate of change.
- With switching observations, variance corresponding to unobserved model can grow without bound, effectively ignoring level-dependent features.
- Without constraints on limits of $\mathbf{g}^{t+1}$, false reject errors may occur when baseline model is a poor fit.

2. Lombard dynamics

$$P(\mathbf{g}^{t+1}|\mathbf{g}^t) \propto \mathcal{N}(\mathbf{g}^{t+1};\ \mathbf{g}^t,\ \Sigma_{RW})\,\mathcal{N}(\mathbf{g}^{t+1};\ \mu_{SNR},\ \Sigma_{SNR})$$

- Add "SNR prior" constraints to random walk model that are independent of the observations.
- Enforces a range on the state variable.
- Enforces a ratio between the speech and noise levels enabling the dynamics to capture the Lombard effect.



References

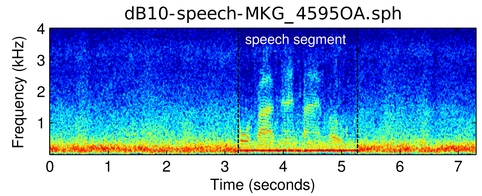
S. Rennie, T. Kristjansson, P. Olsen, and R. Gopinath. Dynamic noise adaptation. In *Proceedings of ICASSP*, 2006.

M. Fujimoto and K. Ishizuka. Noise robust voice activity detection based on switching Kalman filter. In *Proceedings of Interspeech*, pages 2933– 2936, 2007.

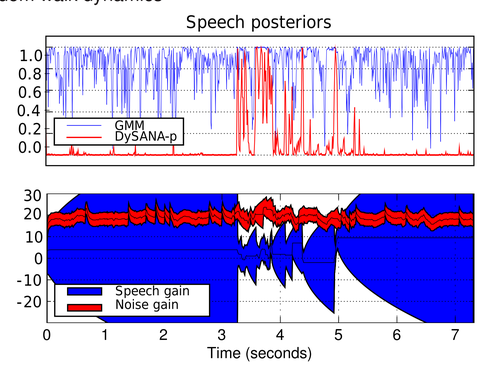
Speech processing, transmission and quality aspects (STQ), distributed speech recognition, advanced front-end feature extraction algorithm, compression algorithms, January 2007. ETSI standard document. ETSI ES 202 050 V1.1.5.

ANSI-C code for the adaptive multi rate (AMR) speech codec, June 2007. 3GPP standard document. 3GPP TS 26.073 V7.0.0.

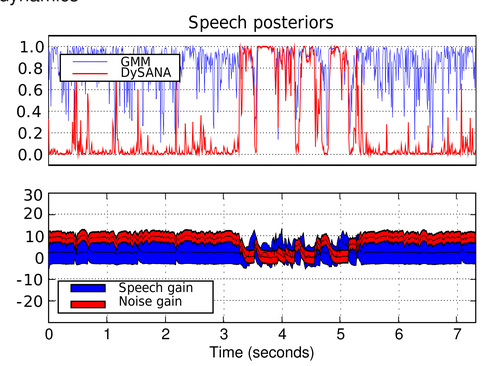
4. Examples



Random walk dynamics



Full dynamics



5. Evaluation

- AURORA-2 speech mixed with DNA database of car noise (Rennie et al., 2006).
- Utterances generated to mimic interactions with a dialog system.
- Silence periods before and after speech segment, 8% contain no speech.
- Passed through AMR codec to simulate cell phone channel.
- 32 component speech/nonspeech GMMs trained on clean data collected from the Goog411 dialog system.
- Evaluated 2 variants of the algorithm
  - DySANA-p - DySANA without prior constraints (i.e. random walk dynamics)
  - DySANA - full system with Lombard dynamic distribution
- Compare to baseline adaptive energy threshold VAD, unadapted GMM classifier, ETSI AFE VAD, and a similar switching Kalman filter system based on parallel model combination (Fujimoto and Ishizuka, 2007).
- Used multicondition AURORA-2 HTK recognizer trained over AMR coded speech.

Table: Word error rate as a function of SNR

System	0	5	10	15	20
No VAD	106.5	97.8	81.7	70.1	63.5
ETSI AFE	93.7	87.5	78.7	59.6	57.5
Energy	106.5	96.5	76.7	56.4	30.0
GMM	79.7	63.4	35.2	22.2	11.5
SKF	78.8	51.6	27.8	17.2	8.7
DySANA-p	**64.6**	47.3	27.7	14.6	7.8
DySANA	74.2	**46.8**	**23.9**	**13.5**	**6.2**

6. Discussion

- GMM endpointer performs very poorly under noisiest conditions because the baseline model is a poor fit for the data.
- DySANA performs best under moderate noise.
- At 0dB SNR DySANA is unable to track extremely high noise levels leading to false accept errors.
- At high SNRs DySANA-p makes more false reject errors than DySANA leading to decreased performance.

Interspeech 2008, 22-26 September 2008, Brisbane, Australia