AN EXPLORATION ON INFLUENCE FACTORS OF VAD'S PERFORMANCE IN SPEAKER RECOGNITION

Cheng Gong, CSLT

2013/04/15

Outline

- Introduction
- Analysis about influence factors of VAD's performance
- Experimental results and analysis
- Conclusions

Introduction

- Voice activity detection (VAD)
  - A method for detecting periods of speech in observed signals
  - VAD technique is particularly important and widely used in both automatic speech recognition and speaker recognition
  - Two parts of VAD process:
    - Acoustic feature extraction
    - Decision mechanism
- Currently used VAD methods:
  - Short-term signal energy, zero-crossing rate
  - Speech/noise spectral characteristics based methods:
    - MFCCs, LTSE, LSF, MMSE, etc.
  - Periodic feature based methods:
    - ACF, F0, etc.
  - ……

Introduction

- Difficulties of VAD:
  - Determine end-points accurately
  - Be robust to noise, especially to non-stationary noise
- Basic principle of choosing end-points:
  - Speech recognition: integrity of the speech contents
  - Speaker recognition: typicality of the speaker characteristics

To get a better result, VAD method in speaker recognition may be different from which in speech recognition

Phonation Types

- Voiced sound
  - Glottis excitation + Vocal Tract response
  - Quasi-periodic signal
  - All simple/compound vowels and 4 initial consonants (m, n, l, r) in mandarin are voiced sound
- Unvoiced sound
  - No vocal cord vibration
  - Non-periodic signal
  - Plosive/affricate/fricative, aspirated/unaspirated
  - The other initial consonants in mandarin are unvoiced sound

Phonation Types' Influence

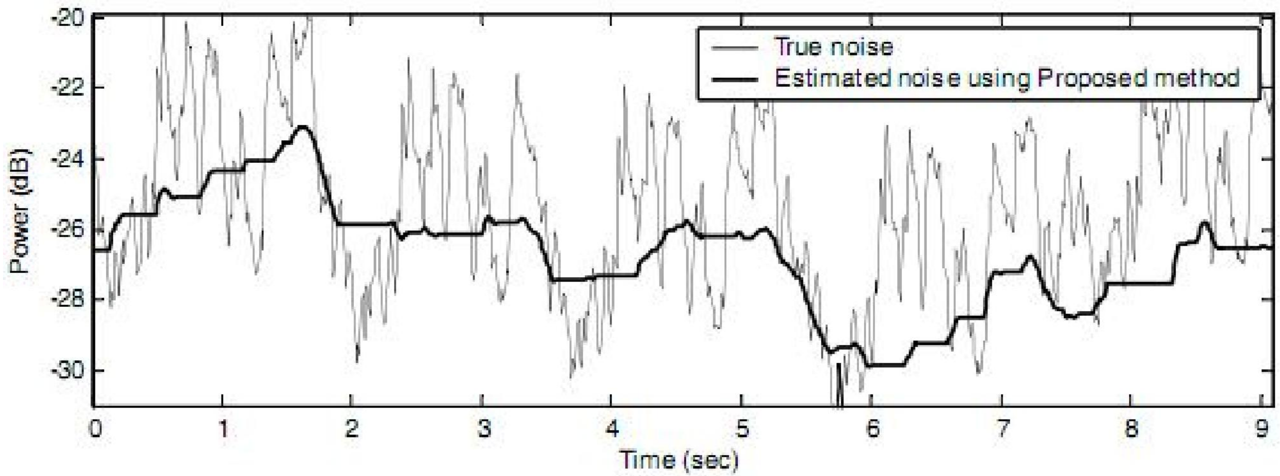
- Research Assumption
  - Phonation types' distinction may lead to different contribution to speaker verification results
- Research Procedure
  - To segment the speech signals based on phonation types (Using HVITE tools)
  - To splice the speech according to the rules of classification by person
    - Silence segments
    - Voiced sound segments
    - Unvoiced sound segments
  - To extract features (MFCC), train models and test on the speaker verification system, compare and analyse the results.

SNR's Influence

- Research Assumption
  - Noise in the speech doesn't reflect the speaker's characteristics, so the parts which has low SNR may lead to a high EER
- Research Procedure
  - To estimate noise power spectrum of each speech signal
  - To Calculate SNR of each frame
  - To splice the speech based on the SNR level
    - 'Clean'~20dB, 20dB~15dB, 15dB~10dB, 10dB~ 5dB, 5dB~
  - To extract features (MFCC), train models and test on the speaker verification system, compare and analyse the results.

Noise Estimation Algorithm

- Analysis object
  - Additive noise in the speech (stable/unstable)
- Destination
  - To obtain a noise power spectrum estimation from noisy speech
- Implement method
  - Combination of minimum statistics, continuous spectral minimum & minima controlled recursive algorithm



Noise Estimation Algorithm

1. Compute smooth speech power spectrum $P(\lambda,k)$
2. Find the local minimum of noisy speech $P_{min}(\lambda,k)$
3. Compute ratio $S_r(\lambda,k)$ of smoothed speech power spectrum to its local minimum
4. Calculate speech presence probability $p(\lambda,k)$ using first-order recursion
5. Compute time-frequency dependent smoothing factors $\alpha_s(\lambda,k)$
6. Update noise estimate $D(\lambda,k)$ using time-frequency dependent smoothing factors $\alpha_s(\lambda,k)$

Database

- CCB database
  - Recorded in clean environments using telephone channel
  - Sampling rate: 8kHz
  - Training utterance length: 39s~75s
  - Testing utterance length: 11s~44s

Channel	**Training**		**True Speaker**		**Impostor**	
Telephone	M	F	M	F	M	F
	50	50	150	150	1000	1000
	100		300		2000	

Experimental Conditions

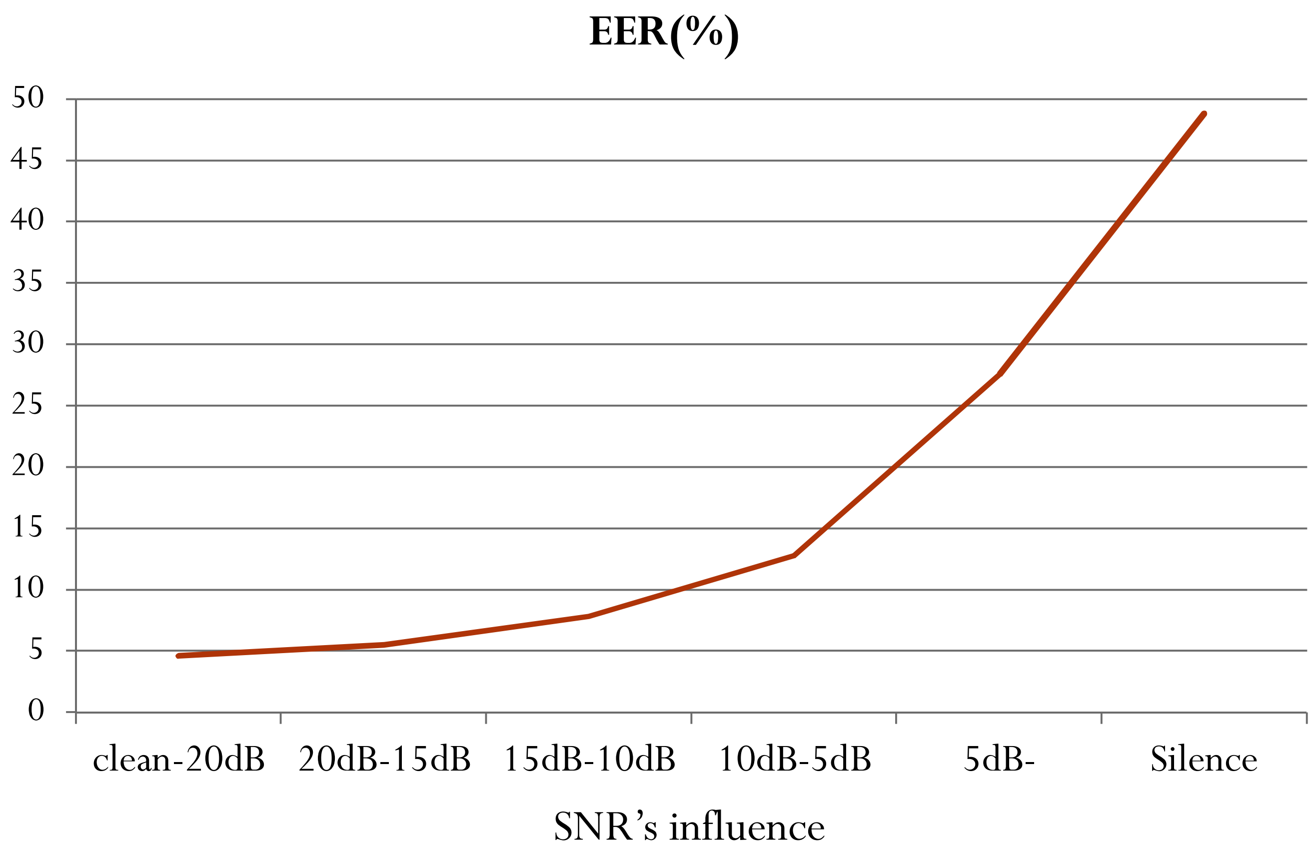
- Feature:
  - Mel-frequency cepstral coefficients (MFCC)
  - 16-orders with energy, without delta
  - 32 Mel filter banks
- Model:
  - GMM-UBM
  - 1024 mixtures

Results and Analysis

Gender	EER(%)			
	Voiced Sound	Unvoiced Sound	Silence	Baseline
M	7.65	42.49	48.74	8.17
F	8.13	42.17	49.22	8.53
M+F	5.89	41.87	49.12	7.44

Phonation types' influence

Results and Analysis



If the segments (SNR<5dB) are removed, the EER = 5.09% (the baseline EER = 7.16%)

Results and Analysis

- Add white noise with different SNRs to the speech, the table below shows the EERs when removing the segments whose SNR<5dB:

SNR	EER(%)	
	Improved	Baseline
clean	5.09	7.16
20dB	6.46	8.76
15dB	8.31	10.89
10dB	11.85	15.24
5dB	16.65	21.46

Conclusion

- Unvoiced sounds don’t contribute much to speaker verification results. The speech with voiced sounds only can get better results.
- The EER is related to SNR of the speech directly, if we remove some segments whose SNR is very low, the results will get much better. This method has remarkable effects on noisy speech.

References

1. Ishizuka,K., Nakatani, T., Fujimoto,M., Miyazaki,N., 2010. Noise robust voice activity detection based on periodic to aperiodic component ratio. Speech Commun. 52,41-60
2. Le Bouquin-Jeannes,R., Faucon,G., 1995. Study of voice activity detector and its influence on a noise reduction system. Speech Commun. 16,245–254
3. Karray,L., Martin,A., 2003. Towards improving speech detection robustness for speech recognition in adverse conditions. Speech Commun.40,261–276
4. Rabiner,L.R., Sambur,M.R., 1975. An algorithm for determining the end points of isolated utterances. BellSyst.Tech.J.54,297–315
5. Kristjansson,T., Deligne,S., Olsen,P., 2005. Voicing features for robust speech detection. In:Proc.Interspeech,pp.369–372
6. Ramirez,J., Segura,J.C., Benitez,C., De la Torre,A., Rubio,A., 2004. Efficient voice activity detection algorithms using long-term speech information. SpeechCommun. 42,271–287
7. ITU-T Recommendation G.729 Annex B, 1996. A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70
8. Ephraim,Y., Malah,D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoustic Speech Signal Process. ASSP-32,1109–1121
9. Rabiner,L.R., Sambur,M.R., 1975. An algorithm for determining the end points of isolated utterances. BellSyst.Tech.J.54,297–315
10. Ahmadi,S., Spanias,A.S., 1999. Cepstrum-based pitch detection using a new statistical V/UV classification algorithm. IEEETrans. on Speech and Audio Process.7,333-338

References

11. Rangachari,S., Loizou,P.C., 2006. A noise-estimation algorithm for highly non-stationary environments. Speech Communication 48 220-231
12. Martin,R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. 9 (5), 504–512
13. Cohen,I., 2002. Noise estimation by minima controlled recursive averaging for robust speech enhancement. IEEE Signal Process. Lett. 9 (1), 12–15
14. Doblinger,G., 1995. Computationally efficient speech enhancement by spectral minima tracking in subbands. Proc. Eurospeech 2, 1513–1516

Thank you!