Selection of optimal vocal tract regions using real-time magnetic resonance imaging for robust voice activity detection

Abhay Prasad[1], *Prasanta Kumar Ghosh*[2], *Shrikanth S Narayanan*[3]

[1]Electronics and Communication Engineering, Manipal Institute of Technology, Manipal, India
[2]Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India
[3]Electrical Engineering, University of Southern California (USC), Los Angeles-90089, U.S.A.

abhayprasad.337@gmail.com, prasantg@ee.iisc.ernet.in, shri@sipi.usc.edu

Abstract

Real time magnetic resonance imaging (rtMRI) enables direct video capture of the moving vocal tract concurrent with audio signal providing valuable data for speech research. We consider a multimodal approach to voice activity detection (VAD) in the rtMRI recording that uses audio signal as well as MRI image sequence. The degraded quality of the audio recorded in the scanner motivates this multimodal scheme for robust VAD. Optimal regions in the MRI image are selected for performing VAD with a novel algorithm. VAD experiments using rtMRI data of two male and two female subjects show that VAD performance using optimally selected regions from MRI images is comparable to that using only audio signal. The optimal regions turn out to be parts of jaw, velum, glottis and lips. VAD performance using audio signal and MRI image sequence together is found to be significantly better ($\sim$14% absolute improvement in VAD accuracy) than that using the audio only when the audio is contaminated with additive noise at low SNR.

Index Terms: voice activity detection, vocal tract imaging, optimal region selection.

1. Introduction

Voice activity detection (VAD) is the task of identifying which regions in a recorded audio signal belongs to the spoken utterances. Thus the challenge in VAD lies in distinguishing speech from non-speech regions in a given audio stream. The non-speech regions could be either silence or any other noise signals. In this work, we focus on VAD of recording from real time Magnetic Resonance Imaging (rtMRI) [1] which simultaneously records the speech signal and the movement of various articulators in the subject's midsagittal plane while the subject is speaking. Such data have opened up new possibilities for speech science research [2, 3, 4, 5, 6]. However, the audio recorded in rtMRI suffers from scanner noise (which includes gradient coil noise, collant pump noise and fan noise); this renders analysis and modeling of speech production challenging. Hence, we consider a multimodal approach to VAD that leverages both the audio and video information of speech articulation available in the rtMRI data.

A typical VAD system works only with the audio signal. Many existing algorithms for audio-only VAD use features that depend on energy [7, 8, 9], zero-crossing rate (ZCR) [10], wavelet transform coefficients [11], cepstral features [12] or their combinations [8, 13]. There are also approach to VAD based on short-time neg-entropy measure [14], long-term spectral divergence [15] and signal variability [16]. In contrast to audio-only VAD, in this work, we would like to examine how the MRI video of articulatory dynamics (by itself and jointly with audio features) can be used to design a robust VAD system for rtMRI recordings, a novel source of data for speech research. It is important to note that the MRI video of the upper airways provides a rich and detailed view of the vocal tract dynamics involved in speech production. Thus, the changes in the vocal tract shape would be more while the subject is speaking compared to when the subject is not speaking. We would like to utilize the information about vocal tract shape change from the MRI video for the VAD task and compare it with the VAD performance obtained from only audio. Since rtMRI video quality remains unaltered in the presence of audio noise, we would also like to investigate if MRI video can be used to design a robust VAD by using the audio and MRI video together particularly when the recorded audio is contaminated by additive noise. Such a multi-modal VAD scheme could be useful for automatically identifying speech segments from noisy rtMRI recordings.

In rtMRI, the upper airway of a subject is imaged; this includes subject's nose, upper palate in addition to the vocal tract region starting from the lips to the glottis. MRI video frames also have regions that are outside the subject's face in the midsagittal plane. Thus, while MRI images contain vocal tract shaping information, automatically finding out regions in MRI images that capture maximal information about speech activity remains to be investigated. It should also be noted that subjects during rtMRI recording could make involuntary movements of the articulators which may be related to non-speech activity such as swallowing, or getting ready to speak [17, 18]. Thus one needs to find regions which can distinguish articulatory movements due to speech from those due to non-speech activities.

In this paper, we propose a region selection algorithm which optimally combines image blocks from different portions of the MRI video frames so that the VAD performance is maximized. Experiments using rtMRI data from multiple subjects demonstrate that the VAD performance using selected MRI image regions is similar to that using only audio signal. The number of pixels in the optimally selected regions turn out to be less than 1% pixels of the entire MRI image for all subjects. Importantly, we also find that combining audio and optimally selected MRI regions yields better VAD performance than when using them separately, particularly at low SNRs. We begin with the description of the rtMRI dataset used for the present study.

2. rtMRI database

We have used a multimodal real-time MRI articulatory corpus [1] for the present study. The rtMRI corpus consists of simultaneous recording of speech and articulatory dynamics in the midsagittal plane acquired from two male (M1 and M2) and two female (F1 and F2) speakers of American English. Speak-

Work supported by Department of Science and Technology (DST), Govt. of India.

ers' upper airways are imaged in the midsagittal plane while they read the same 460 sentence corpus used in the MOCHA-TIMIT corpus [19]. The image resolution in the sagittal plane is 68×68 pixels. Image data has a frame rate of 23.18 frames/sec. Audio data is simultaneously recorded at a sampling frequency of 20kHz inside the MRI scanner while subjects are imaged. A specially designed noise cancellation technique is used to remove scanner noise from the recorded audio [20]. The total duration of the recordings are 39.05, 38.07, 38.19, and 37.99 minutes for M1, M2, F1, and F2 respectively. During rtMRI recording, the 460 sentences are split into 92 groups each with 5 sentences. Thus rtMRI corpus provides 92 video files for each of the four subjects. The subject read each TIMIT sentence with pause both before and after the sentence. The average duration of inter-utterance pauses are 1.25, 1.45, 1.36, and 1.75 seconds for the four subjects. Since the amount of inter-utterance pause varies from one subject to another, the speech data of 460 sentences would correspond to different percentages of the entire recording for different subjects. For example, only 75.45% of the entire recording for M1 corresponds to speech. The fractions are 70.61%, 72.61%, and 64.58% for M2, F1, and F2 respectively.

3. Optimal region selection in MRI image

Let $M_n(i,j)$, $1 \le i \le I$, $1 \le j \le J$ denotes the image of the n-th frame in the MRI video, where $M_n(i,j)$ is the intensity of the (i,j)-th block. Consider there are N such image frames (i.e., $1 \le n \le N$) in the training dataset where each frame is manually annotated with a label which indicates whether the respective frame belongs to speech or not. We consider $P \times P$ non-overlapping blocks in the MRI image. Thus, the number of blocks over the entire MRI image is $\lfloor \frac{I}{P} \rfloor \times \lfloor \frac{J}{P} \rfloor$ [1]. Let $B(k,l)$ denotes a block located at the (k,l)-th pixel, where $1 \le k \le K = \lfloor \frac{I}{P} \rfloor$ and $1 \le l \le L = \lfloor \frac{J}{P} \rfloor$. We compute the average of all the pixel values within each block as follows:

$$S_n(k,l) = \frac{1}{P^2} \sum \sum_{i,j \in B(k,l)} M_n(i,j) \quad (1)$$

This is obtained by filtering an MRI image using a moving average filter with an impulse response of size $P \times P$. We assume that during speech activity the blocks on the vocal tract regions would show a wider variability in $S_n(k,l)$ over several consecutive frames. To capture this variability we compute the standard deviation of $S_n(k,l)$ at the n-th frame using a window of length $2D+1$ as follows:

$$\gamma_n(k,l) = \sqrt{\frac{1}{2D+1} \sum_{m=n-D}^{m=n+D} \big(S_m(k,l) - \bar{S}_n(k,l)\big)^2}, \quad (2)$$

where $\bar{S}_n(k,l) = \frac{1}{2D+1} \sum_{m=n-D}^{m=n+D} S_m(k,l)$. Thus, we hypothesize that $\gamma_n(k,l)$ would be high at the n-th frame if the speech is present at the n-th frame and if the block at (k,l)-th pixel reflects the articulatory motion due to speech activity. On the other hand, $\gamma_n(k,l)$ would be low if either speech is not present at the n-th frame or the block at the (k,l)-th pixel does not reflect any articulatory motion due to speech. We quantify the potential of each block for VAD task by computing the receiver operating characteristic (ROC) curve using $\gamma_n(k,l)$ and the ground truth speech activity label of all N frames available from manual annotation. In order to minimize both false alarms and true rejections, we compute the equal error rate (EER) as the measure of the VAD performance by a block. Thus we obtain EER matrix denoted by $E(k,l)$, $1 \le k \le K, 1 \le l \le L$. The lower the $E(k,l)$, the better is the block at (k,l)-th pixel for VAD. Using N = 204921, 205400, 211432, 205768 frames for the four subjects in the rtMRI corpus, the obtained EER matrices are shown in Fig. 1. P and D are set to 2 and 7 respectively. It is clear that, in general, blocks falling on lips, jaw, tongue surface, velum and glottis result in low EER. It is expected since these articulators are typically used for speech production. However, the EER obtained by each of these articulators are different for different subjects, e.g., for F1 and F2 the jaw regions have EER close 0.2 but those for M1 and M2 are higher than 0.2. Similarly the glottis regions for male subjects resulted in least EERs while that is not the case with female subjects.

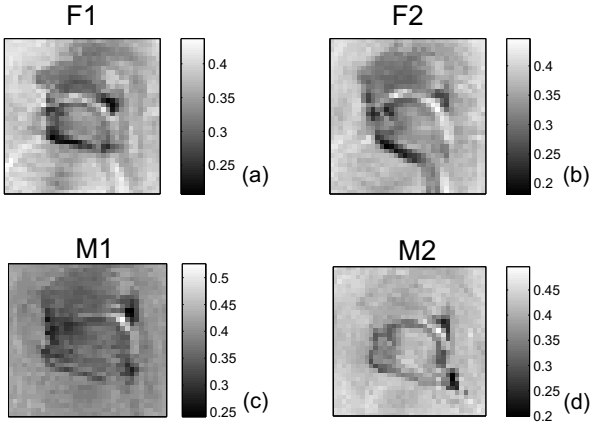

Figure 1: *Illustration of EER obtained by different blocks for four subjects in the rtMRI corpus.*

Let us denote all KL blocks using one index κ where there exist a k and an l so that $\kappa = (k-1)J + l$. Blocks with this new index are denoted by B^κ, $\kappa = 1, \cdots, KL$. The corresponding EER is denoted by E^κ. Let E_s^κ be the sorted EERs in ascending order. The corresponding blocks are denoted by B_s^κ. We perform a forward block selection method to find optimal regions in the MRI image. Regions are formed by taking union of the blocks from B_s^κ corresponding to the low EERs. This begins with B_s^1. The next R blocks in terms of the least EER (i.e.,B_s^r, $1 \le r \le R$) are considered. The region formed using these R blocks (denoted by B_s^{1-R}) is used to compute the average pixel intensity S_n^{1-R}, which is further used to compute the variability measure γ_n^{1-R} in a manner similar to eqn (2) as follows:

$$B_s^{1-R} = \bigcup_{1 \le r \le R} B_s^r, \quad S_n^{1-R} = \frac{1}{P^2 R} \sum \sum_{i,j \in B_s^{1-R}} M_n(i,j)$$

$$\gamma_n^{1-R} = \sqrt{\frac{1}{2D+1} \sum_{m=n-D}^{m=n+D} \big(S_m^{1-R} - \bar{S}_n^{1-R}\big)^2}, \quad (3)$$

where, $\bar{S}_n^{1-R} = \frac{1}{2D+1} \sum_{m=n-D}^{m=n+D} S_m^{1-R}$. The EER ($E^{1-R}$) is computed using γ_n^{1-R}, which measures the potential of the regions obtained by combining R blocks with least EERs. Since each of these blocks have low EER, combining them could reduce the EER further. However, if R is too large, the combined region may contain blocks which are not from the vocal tract areas in the MRI image. Therefore, we examine the E^{1-R} for different values of R and choose the $R = R^\star$ so that $E^{1-R^\star}$ attains the minimum value. The region obtained by combining these $R^\star$ blocks is declared to be the optimally selected region. The algorithm for selecting optimal region is summarized in Algorithm 1.

[1] $\lfloor x \rfloor$ is an integer closest to x and less than x.

Algorithm 1 Selection of optimal region $B_s^{1-R^\star}$ in the MRI image for VAD

Inputs: B_s^κ, $1 \leq \kappa \leq KL$, all blocks in the MRI image sorted according to the their EERs in ascending order.
Initialization: R=1.
while $R \leq KL$ **do**
  Compute combined region $B_s^{1-R} = \bigcup_{1 \leq r \leq R} B_s^r$
  Compute S_n^{1-R} and γ_n^{1-R} using eqn (3)
  Compute EER E^{1-R} using γ_n^{1-R} and the ground truth speech activity label
  $R \leftarrow R + 1$
end while
$R^\star = \min_R E^{1-R}$, $B_s^{1-R^\star} = \bigcup_{1 \leq r \leq R^\star} B_s^r$.
Return $B_s^{1-R^\star}$.

4. Experiments and results

4.1. Experimental setup

The VAD experiment is performed separately for each subject of the rtMRI corpus. This is done to examine how the optimal regions in the MRI image for VAD varies across subjects. For each subject. 92 recordings are used for VAD experiments in a 9-fold cross-validation setup [2]. Eight folds are used for training and the remaining fold is used as the test set. Audio stream and the MRI image sequences are extracted from each of the rtMRI videos. Three different VAD schemes are used for comparison: 1) audio-only VAD (denoted by VAD_audio), 2) VAD using optimal regions from MRI images (VAD_rtMRI) and 3) VAD using both audio and optimal regions from MRI images (VAD_audio_rtMRI). Since a major portion of the recorded signal is speech, we use a baseline scheme (denoted by VAD_baseline) which declares the entire test utterance to be speech. The VAD decision is taken every 10 msec using each of these VAD schemes. Details of the three VAD schemes are described below.

VAD_audio

The speech and non speech segments in the audio are detected using a VAD algorithm based on Long Term Signal Variability (LTSV) [16], which measures the degree of non-stationarity in the audio signal. It has been shown [16] that the LTSV based VAD scheme performs better at different SNRs and different noises over standardized VAD schemes such as AMR VADs option 1 and 2 [21] and ITU G.729 AnnexB VAD [22]. Hence we use the LTSV based VAD scheme for comparison in this study. To obtain the best VAD_audio performance at different SNRs we have trained the LTSV based VAD at each SNR separately where the parameters of LTSV are optimized on the training set.

VAD_rtMRI

The VAD_rtMRI works based on the variability measure $\gamma_n^{1-R^\star}$ over the optimally selected regions from the MRI image sequence in the training data. The threshold corresponding to minimum EER on the training data is used for VAD on the test set. The variability measure on an image frame from the test set is computed using optimally selected regions from the training data. If this variability is greater than the threshold, the respective frame is detected as speech. Performance of VAD_rtMRI does not change at different SNRs unlike audio because MRI images remain unaffected due to noise in the audio domain. It is important to note that the MRI frame rate is 23.18 frames/sec but the VAD is done at a frame rate of 100Hz. To compensate for this mismatch in frame rates, we upsample the MRI video to 100Hz by linear interpolation. We have experimented with P=2, 4 and D=1, 3, 5, 7. P=2 and D=7 yield the best VAD performance on the training set for all folds of different subjects. Hence performance of VAD_rtMRI are reported with this choice of parameter values.

VAD_audio_rtMRI

LTSV from audio and $\gamma_n^{1-R^\star}$ from MRI image are used to form a 2-dimensional feature for VAD using VAD_audio_rtMRI. The speech non-speech binary classification is done using a neural network with 1 hidden layer. The hidden layer consists of 10 neurons, with the tan sigmoid activation function. The output layer consists of a single neuron having tan sigmoid activation function as well. The neural network is trained with 2-dimensional features of all frames in the training corpus; training is done using resilient back-propagation algorithm [23]. Unlike VAD_audio, the neural network is not trained at each SNR condition separately. It is trained only for the clean case and used directly on the feature vectors at all other SNRs.

We report the VAD performance using VAD_audio, VAD_rtMRI, and VAD_audio_rtMRI when the recorded audio signal is contaminated with additive white and babble noise at different SNRs: -20dB, -10dB, 0dB, 10dB, 20dB and clean. Noise samples are taken from NOISEX-92 corpus [24]. This is done to examine whether the information about speech activity from MRI images could complement to that from audio in cases when audio quality is degraded with additive noise. It should be noted that a hangover scheme [25] is typically used to postprocess the VAD decisions. However, we have not found any significant improvement due to a hangover scheme. Hence, no hangover scheme is used for all results reported in the work. The VAD performance is reported following the work by Freeman et al [26] and Beritelli et al [27]. Five different parameters reflecting VAD performance are reported, namely, accuracy of VAD (ACU), front end clipping (FEC), mid speech clipping (MSC), carry over (OVER), noise detected as speech (NDS). FEC and MSC reflect true rejection, while NDS and OVER reflect false acceptance. Thus low values of FEC, MSC, NDS, OVER and high value of ACU would indicate a good VAD performance.

4.2. Results and discussions

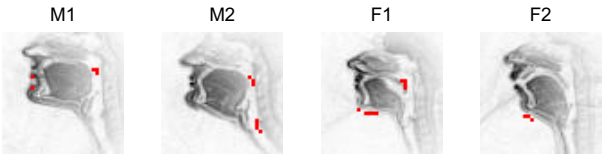

Figure 2: *Optimal regions (marked in red) in the MRI images for VAD task for four different subjects.*

Fig. 2 illustrates the optimal regions selected using the proposed algorithm for all four subjects in the rtMRI corpus. These optimal regions are derived using the training data from the first fold for all the subjects. The optimal regions do not change much across different folds of a single subject. The pixels in the selected regions are marked with red. The number of pixels corresponding to the optimal regions are 20, 28, 32, and 12 for M1, M2, F1, and F2 respectively. These correspond to 0.43%, 0.61%, 0.69%, and 0.26% of the total number of pixels in an MRI image. This suggests that the speech activity information is captured by a small fraction of the pixel intensities in an MRI image. In spite of variation in the vocal tract morphology of different subjects, these optimal regions coincide with jaw, velum, lower lip and glottis for different subjects. Among four subjects, a portion of velum is selected as the optimal region for the two male and one female subject. This indicates

[2] one fold contains 12 recordings and each of remaining eight folds contains 10 recordings

that pixel intensity variation in the velum region could provide a robust cue about speech activity. This is because, only during speaking, the velum is lifted to close the nasal path (except for nasal sounds). Such motion of velum does not happen for other involuntary movement such as swallowing or rest unlike other articulators like lips, tongue. This could be the reason why the tongue is not selected at all as the optimal region for any of the four subjects. The glotti region is found to belong to the optimally selected regions only for one male subject (M2). Different parts of the jaw belong to the optimal regions for three different subjects suggesting movement of jaw could also be a robust indicator of speech activity.

Fig. 3 shows the VAD accuracies obtained by VAD_audio, VAD_rtMRI, VAD_audio_rtMRI, and VAD_baseline separately for each subject at different SNRs with additive white and babble noise. It is clear that the accuracy by VAD_rtMRI is significantly greater than that using VAD_baseline for all subjects: 2.1%, 11.2%, 8.7%, and 16.3% absolute improvement for M1, M2, F1, and F2 respectively. On an average, the accuracy of VAD_audio is greater than that of VAD_rtMRI when the VAD is performed on the clean audio signal. However the VAD_audio performance gradually falls as SNR decreases. The accuracy of VAD_audio drops below VAD_baseline at -10dB SNR in the case of additive white noise and at 0dB in the case of additive babble noise. Since white noise is more stationary than babble noise, the LTSV measure [16] separates white noise from non-stationary speech with greater accuracy than babble noise from speech. Thus, at a given SNR, VAD_audio performs better in the case of white noise than babble noise. However, at -20dB the accuracy of VAD_audio drops well below VAD_baseline for both types of noises. The accuracy of VAD_audio_rtMRI is similar to that of VAD_audio in the clean condition. However, as SNR decreases the accuracy of VAD_audio_rtMRI remains close to that of VAD_rtMRI. Even at -20dB SNR condition, the accuracy of VAD_audio_rtMRI is significantly greater than that of VAD_baseline except for M1 and M2 with additive white noise. It should be noted that in the case of VAD_audio_rtMRI, the neural network classifier is trained only in the clean condition and directly is used for VAD in different SNR conditions. Superior performance of the VAD_audio_rtMRI over VAD_audio indicates that, at low SNR, optimally selected regions in the MRI image plays a crucial role in compensating for the poor audio quality for VAD task.

For a detailed comparison across VAD_audio, VAD_rtMRI, VAD_audio_rtMRI, and VAD_baseline, different VAD performance measures FEC, MSC, NDS, OVER, ACC are reported at three different SNRs (-20dB, 0dB, clean) averaged over all subjects and noise types in Table 1. It is clear that the ACC of VAD_audio_rtMRI is better than VAD_rtMRI and similar to VAD_audio in the clean condition. However at 0dB, ACC of VAD_audio_rtMRI becomes significantly better than both VAD_audio and VAD_rtMRI. At -20dB, ACC of VAD_audio_rtMRI becomes similar to that of VAD_rtMRI and significantly better than VAD_audio. This suggestz that although rtMRI video may not provide information complementary to audio in clean condition, rtMRI video compensates for audio quality degradation for VAD at low SNR. It is interesting to note that VAD_rtMRI has the least value of FEC and OVER. However, it results in higher values of NDS suggesting that many MRI frames were detected as a speech frame although subject was not speaking. This false alarm could happen due to articulatory movements of non-speech activities.

5. Conclusions

We propose an algorithm for multimodal VAD which uses both audio as well as MRI image sequence that captures the vocal tract dynamics. We find optimal regions in the MRI image that results in best VAD performance. Using audio features and features derived from the optimal regions in MRI images, we show that the multimodal VAD is superior to audio-only based VAD particularly when the audio is degraded with additive noise at low SNR. The proposed region selection approach could also be useful for deriving features in forced alignment or other data-driven articulatory modeling. However, the sequential nature of the region selection algorithm could result in a sub-optimal region. In fact, the selected region may even fall outside the vocal tract area. But we have not encountered any such in our experiemnts. A region growing algorithm can instead be used to avoid any constraint due to block size and shape. The proposed algorithm can also be evaluated for audio contaminated with real MRI scanner noise to provide more insight into the performance of the algorithm in real time recording scenario. These are parts of our future work.

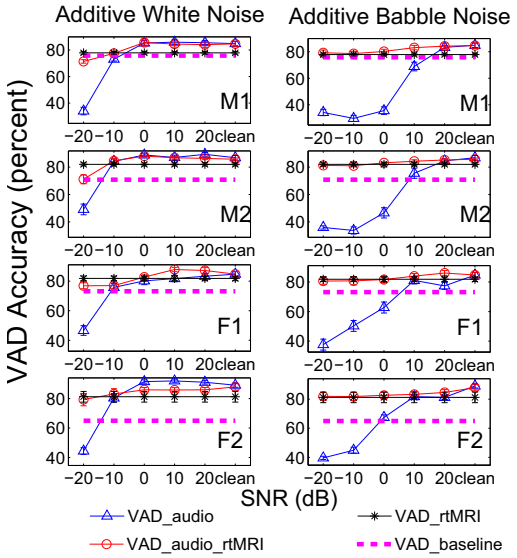

Figure 3: *VAD accuracies averaged over all folds of each subject using VAD_audio, VAD_rtMRI, VAD_audio_rtMRI, and VAD_baseline schemes at different SNR and noise conditions. Errorbars indicate the standard deviation of the VAD accuracy over all folds.*

SNR		VAD schemes audio	rtMRI	audio_rtMRI
-20dB	FEC	0.34(0.05)	0.02(0.02)	0.02(0.01)
	MSC	0.78(0.02)	0.19(0.02)	0.2(0.04)
	NDS	0.15(0.03)	0.19(0.06)	0.23(0.06)
	OVER	0.04(0.02)	0.01(0.01)	0.03(0.01)
	ACC	0.40(0.02)	0.81(0.02)	0.78(0.02)
0dB	FEC	0.16(0.02)	0.02(0.02)	0.02(0.01)
	MSC	0.40(0.03)	0.19(0.04)	0.11(0.03)
	NDS	0.04(0.02)	0.19(0.07)	0.28(0.06)
	OVER	0.02(0.01)	0.01(0.01)	0.06(0.03)
	ACC	0.69(0.02)	0.81(0.02)	0.84(0.01)
Clean	FEC	0.09(0.01)	0.02(0.02)	0.05(0.01)
	MSC	0.14(0.02)	0.19(0.04)	0.11(0.02)
	NDS	0.10(0.04)	0.19(0.07)	0.22(0.05)
	OVER	0.07(0.03)	0.01(0.01)	0.06(0.03)
	ACC	0.86(0.02)	0.81(0.02)	0.86(0.02)

Table 1: *VAD performance measures of different VAD schemes at three different SNRs averaged over all subjects and noise types. ACC for VAD_baseline is 0.71(0.02).*

6. References

[1] S. Narayanan, E. Bresch, P. K. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. C. Lammert, M. I. Proctor, V. Ramanarayanan, and Y. Zhu, "A multimodal real-time mri articulatory corpus for speech research," *Proceedings of Interspeech*, pp. 837–840, 2011.

[2] C. Hagedorn, M. Proctor, and L. Goldstein, "Automatic analysis of geminate consonant articulation using real-timemagnetic resonance imaging," *in Proc. 9th ISSP*, 2011.

[3] A. Lammert, M. Proctor, and S. Narayanan, "Morphological variation in the adult vocal tract: A study using rtmri," *in Proc. 9th ISSP*, 2011.

[4] M. I. Proctor, N. Katsamanis, L. Goldstein, C. Hagedorn, A. Lammert, and S. Narayanan, "Direct estimation of articulatory dynamics from real-time magnetic resonance image sequences," *in Proc. Interspeech*, 2011.

[5] E. Bresch, A. Katsamanis, L. Goldstein, and S. Narayanan, "Statistical multi-stream modeling of real-time mri articulatory speech data," *in Proc. Interspeech*, 2010.

[6] A. Katsamanis, E. Bresch, V. Ramanarayanan, and S. Narayanan, "Validating rt-mri based articulatory representations via articulatory recognition," *in Proc. Interspeech*, 2011.

[7] P. S. H. Krishnan and H. A. Murthy, "Voice activity detection using group delay processing on buffered short-term energy," *Proc. of 13th National Conference on Communications*, 2007.

[8] S. A. Soleimani and S. M. Ahadi, "Voice activity detection based on combination of multiple features using linear/kernel discriminant analyses," *Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on*, pp. 1–5, 2008.

[9] G. Evangelopoulos and P. Maragos, "Speech event detection using multiband modulation energy," *Proc. Interspeech*, pp. 685–688, 2005.

[10] B. Kotnik, Z. Kacic, and B. Horvat, "A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction algorithm," *Proc.7th EUROSPEECH*, pp. 197–200, 2001.

[11] Y. C. Lee and S. S. Ahn, "Statistical model-based vad algorithm with wavelet transform," *IEICE Trans. Fundamentals*, vol. E89-A, pp. 1594–1600, 2006.

[12] J. Haigh and J. S. Mason, "A voice activity detector based on cepstral analysis," *Proc. 3rd EUROSPEECH*, pp. 1103–1106, 1993.

[13] S. McClellan and J. D. Gibson, "Variable-rate celp based on subband flatness," *IEEE Trans. Speech Audio Proc.*, vol. 5, pp. 120–130, 1997.

[14] R. Prasad, H. Saruwatari, and K. Shikano, "Noise estimation using negentropy based voice- activity detector," *47th Midwest Symposium on Circuits and Systems*, vol. 2, pp. II–149 – II–152, 2004.

[15] J. Ramirez, J. C. Segura, C. Benitez, A. Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, pp. 271–287, 2004.

[16] P. G. Ghosh, A. Tsiartas, and S. S. Narayanan., "Robust voice activity detection using long-term signal variability," *Audio, Speech, and Language Processing, IEEE Transactions*, pp. 600–613, 2011.

[17] V. Ramanarayanan, L. Goldstein, D. Byrd, and S. Narayanan, "An investigation of articulatory setting using real-time magnetic resonance imaging," *J. Acoust. Soc. Am. 134*, vol. 4, 2013.

[18] V. Ramanarayanan, E. Bresch, D. Byrd, L. Goldstein, and S. Narayanan, "Analysis of pausing behavior in spontaneous speech using real-time magnetic resonance imaging of articulation," *J. Acoust. Soc. Am. Express Letters 126*, vol. 5, pp. 160–165, 2009.

[19] A. A. Wrench and H. J. William, "A multichannel articulatory database and its application for automatic speech recognition," *5th Seminar on Speech Production: Models and Data, Bavaria*, pp. 305–308, 2000.

[20] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during real-time mri scans," *J. Acoust. Soc. Am. 120*, vol. 4, 2006.

[21] "Voice activity detector for adaptive multi-rate (AMR) speech traffic channels. general description (GEM 06.94 version 7.1.1)," *ETSI EN*, vol. 7, 1999.

[22] "Coding of speech and 8 kbit/s using conjugate structure algebraic code - excited linear prediction. annex b: A silence compression scheme for g.729 optimized for terminals conforming to recommend."

[23] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The rprop algorithm in neural networks," *IEEE International Conference on Neural Network*, pp. 586–591, 1993.

[24] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247–251, 1993.

[25] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. on Audio, Speech and Language Proc.*, pp. 412–424, 2006.

[26] D. K. Freeman, C. B. Southcott, I. Boyd, and G. Cosier, "A voice activity detector for pan-european digital cellular mobile telephone service," *Proc. IEEE ICASSP*, vol. 1, pp. 369–372, 1989.

[27] F. Beritelli, S. Casale, and A. Cavallaro, "A robust voice activity detector for wireless communications using soft computing," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 1818–1829, 1998.