Voice activity detection in a regularized reproducing kernel Hilbert space

Xugang Lu [1], *Masashi Unoki* [2], *Ryosuke Isotani* [1], *Hisashi Kawai* [1], *Satoshi Nakamura* [1]

[1]National Institute of Information and Communications Technology, Japan
[2]Japan Advanced Institute of Science and Technology, Japan

Abstract

Voice activity detection (VAD) is used to detect whether the acoustic signal belongs to speech or non-speech clusters based on the statistical distribution of the acoustic features. Traditional VAD algorithms are applied in a linear transformed space without any constraint relating to the special characteristics speech or noise. As a result, the VAD algorithms are not robust to noise interference. Considering that speech is a special type of acoustic signal that only occupies a small fraction of the whole acoustic space, we proposed a new speech feature extraction method by giving constraints on the processing space as a reproducing kernel Hilbert space (RKHS). In the RKHS, we regarded the speech estimation as a functional approximation problem, and estimated the approximation function via a regularized framework in the RKHS. Under this framework, we could incorporate the nonlinear mapping functions in the approximation implicitly via a kernel function. The approximation function could capture the nonlinear and high-order statistical regularities of the speech. Our VAD algorithm is designed on the basis of the power energy in this regularized RKHS. Compared with a baseline and G.729B VAD algorithms, experimental results showed the promising advantages of our proposed algorithm.

Index Terms: Statistical learning, reproducing kernel Hilbert space, voice activity detection

1. Introduction

Voice activity detection (VAD) is an algorithm that is used to detect whether there exists speech events in an acoustic signals. It is very important and widely used in speech communication technologies, for example, speech recognition, speech enhancement, and speech coding [1]. The task can be regarded as a statistical detection problem for speech absence condition H_0 and speech presence condition H_1 as follows:

$$\begin{aligned} H_0 &: y(t) = v(t) \\ H_1 &: y(t) = x(t) + v(t); \end{aligned} \tag{1}$$

where $y(t)$ is the observation signal, $x(t)$ is the speech signal, and $v(t)$ is the non-speech signal (silence or background noise). Based on Eq. (1), the speech and non-speech can be formulated in a statistical inference problem as likelihood ratio test [2]. The decision is made based on the assumption that speech and non-speech signals are different in their statistical distributions. Although the task is simple, it is a difficult problem in adverse environments because the background noise may degrade the statistical properties of the speech signals. Therefore, robust VAD algorithms are required in real applications. The robustness of an VAD algorithm means that the VAD can give decisions on speech and non-speech close to a reference in clean as well as in noisy environments. Generally speaking, for designing a robust VAD algorithm, two aspects must be considered, one is the noise robust speech features, i.e., in which domain the statistical detection is applied as used in Eq. (1). The other is the selection of decision rules, i.e., what kinds of classifiers should be used to discriminate speech from non-speech based on the features [1, 2]. In this study, we mainly focus on the robust feature aspect for VAD.

Several speech features have been used for VAD, for example, the energy level, zero-crossing, pitch, linear prediction coefficient (LPC) feature, and cepstral feature. Most of them can work accurately in clean environments, but fail when the background noise level increases. Recently, noise robust features, for example, long temporal statistical features, periodicity measure, and high-order statistics in the LPC residual space were proposed for VAD in noisy environments [1, 4]. In order to reduce the noise effect, noise reduction algorithms were applied to speech enhancement, and the VAD was a byproduct of the algorithms, for example, spectral subtraction, minimum mean square estimation (MMSE) based noise reduction. During noise reduction, the VAD is used for updating the statistical estimation of noise, and the estimated noise is used for signal to noise ratio (SNR) estimation which is used for updating the VAD. Furthermore, the dynamical state modeling for speech was also used in designing the VAD, for example, the generalized Autoregressive conditional heteroskedasticity (GARCH) model, and switch Kanlman filtering [5, 3].

However, most of the current VAD algorithms are applied in a linear transformed space that extracts the linear statistical average or correlations of the acoustic signals, for example, the energy level based VAD relies on the statistical mean estimation of the waveform (first-order statistical information), and the LPC or power spectrum feature based VAD is based on the linear correlation estimation of the waveform (second-order statistical information). Speech is a special type of acoustic signal, it is produced by the movements of articulatory organs with linguistic structure. Its statistical characteristic is different from that of noise, and only occupies a small fraction of the signal subspace of the whole acoustic space. In traditional processing space (via mapping functions), it is possible that the statistical distributions of the speech and non-speech (or noise) are overlapped since speech and noise may have similar linear or low-order statistical structures. For designing VAD under noisy environments, we must give constraints on the mapping functions to get the subspace in which most of the speech information is kept while the noise information is discarded. This consideration is well fit to the functional approximation and generalization problem in the machine learning theory. In this study, we propose to use regularization theory similarly as used in machine learning field to find mapping functions for VAD. In addition, the mapping functions are chosen in a reproducing kernel Hilbert space (RKHS), which is used to obtain nonlinear and high-order statistical information of the data [6]. Our experimental results showed the effectiveness of the proposed

algorithm.

2. Signal approximation in a reproducing kernel Hilbert space

The estimation of the clean speech signal from the observation signal (speech distorted by exterior noise) can be regarded as a learning problem with statistical inference to estimate a target function or predictive function for new testing samples. The main goal for this problem is to select mapping functions from a possible function sets in a functional space. A good choice of the function should give good estimation or encode most information of clean speech even in adverse noisy conditions. We start to consider this problem by using learning theory. Mathematically, we represent an observation as follows:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i \tag{2}$$

For this observation, we try to approximate or learn the target function $f(\cdot)$ from an observation data set $S = \{(\mathbf{x}_i, y_i); i = 1, ..., l\}$, $\mathbf{x}_i \in R^d$ is a vector, and $y_i \in R$ is the response or label information for classification tasks (we will explain how to construct the data pair $(\mathbf{x}_i, y_i)$ from the noisy observations later). The finding of the function $f(\cdot)$ is an ill-posed problem in statistical learning theory since there are many possibilities for the selection of the mapping functions if there is no constraint on the functional space. In order to make the problem to be well posed, we suppose that the $f(\cdot)$ is in a reproducing kernel Hilbert space with a certain smoothness that can be used to approximate the speech component as follows [7]:

$$\hat{y}_i = f(\mathbf{x}_i) = \mathbf{w}^{\mathrm{T}} \Phi(\mathbf{x}_i) \tag{3}$$

where $\Phi(\cdot)$ is a mapping function that maps a vector to a high dimensional space, and $\mathbf{w}$ is the weighting coefficient vector that uniquely determines the target function $f(\cdot)$. Hence the problem is to find a mapping function $f(\cdot)$ by minimizing an objective function $H(f)$ as follows:

$$\begin{aligned} f^* &= \arg\min_f H(f) \\ H(f) &= \tfrac{1}{l} \sum_{i=1}^{l} (y_i - f(\mathbf{x}_i))^2 + \lambda \|f\|_K^2 . \end{aligned} \tag{4}$$

There are two components in the objective function $H(f)$, i.e., the approximation error and the smoothness of the function $f(\cdot)$. The $\|f\|_K^2$ is the norm of the function in a reproducing kernel Hilbert space corresponding to a kernel matrix K constructed from the training data set via the mapping function $\Phi(\cdot)$. The λ is the regularization parameter to make a trade-off between the approximation error and the smoothness of the function. Based on the representer theorem [6], the solution satisfies:

$$f(\mathbf{x}) = \sum_{i=1}^{l} c_i K(\mathbf{x}, \mathbf{x}_i) \tag{5}$$

In Eq. (5), $K(\cdot, \cdot)$ is the kernel function which creates a Gram matrix K with elements defined as follows:

$$K(\mathbf{x_n}, \mathbf{x_m}) = \Phi(\mathbf{x_n})^{\mathrm{T}} \Phi(\mathbf{x_m}) \tag{6}$$

In real applications, we do not need to know the mapping function $\Phi(\cdot)$ explicitly. We only need to calculate the inner product of the mapped vectors via kernel functions. The kernel functions can be chosen as the Gaussian kernel function, or polynomial function which are widely used in statistical learning field [7].

In Eq. (5), c_i is the coefficient which depends on the training data samples. By using the representer theorem, the coefficient vector can be obtained by solving the problem in Eq. (4) as follows [7]:

$$\mathbf{c} = (K + \lambda l \mathbf{I})^{-1} \mathbf{y}, \tag{7}$$

where $\mathbf{I}$ is the identity matrix, the coefficient vector $\mathbf{c} = [c_1, ..., c_l]^{\mathrm{T}}$, and the observation vector $\mathbf{y} = [y_1, ..., y_l]^{\mathrm{T}}$. For learning a predictive or approximation function in Eq. (4) with observation sequence y_i, we reformulate the data in the form of training data pair $(\mathbf{x}_i, y_i)$ with the input $\mathbf{x}$ formulated as $\mathbf{x}_i = [y_{i-p}, y_{i-p+1}, ..., y_{i-1}]$, where p is the dimension of the data vector. In our study, by implicitly choosing a nonlinear mapping via a kernel function, we can approximate the signal by keeping the nonlinear and high-order statistic information of the signal in a regularized RKHS.

3. Voice activity detection based on the power energy in the reproducing kernel Hilbert space

The power energy of speech signal is often used as one of the most simple features for VAD algorithms. The power energy in the original input space (waveform) for one frame is defined as follows:

$$E_y \triangleq \frac{1}{l} \sum_{i=1}^{l} y_i^2 \tag{8}$$

For a zero mean signal, it is the variance of the signal. For clean speech signal, it works quite well for VAD with energy threshold methods. However, for noisy signal, the noise and speech energies are mixed together, it is difficult to use this energy based method for VAD. Considering that the mapped signal in the RKHS, the speech information is well kept while most of the noise information is discarded (due to the smoothness constraint of the mapping functions), we can apply the simple power energy threshold methods for VAD in the RKHS. From Eq. (3), we can see that the mapped signal is uniquely determined by the coefficient $\mathbf{w}$. The energy is defined as the norm of the mapping function $f(\cdot)$ in the RKHS as follows:

$$E_{RKHS} \triangleq \|f\|_K^2 = \mathbf{w}^{\mathrm{T}} \mathbf{w} = \mathbf{c}^{\mathrm{T}} K \mathbf{c} \tag{9}$$

Based on this definition of the power energy in the RKHS, we can simply design a classifier for VAD. The performance of the VAD is expected to be robust in noisy environments.

4. Evaluations

In this section, we test the performance of our proposed processing for VAD, and compare the performance with those of a baseline and the standard G.729B VAD algorithms [10]. In our proposed algorithm, the polynomial function with degree two is used for the kernel function in Eq. (6). The regularization parameter λ in Eq. (4) is set to be 0.5. In construction of the Gram matrix K, we first make frame-based data vectors as segments with 32 ms frame length, and 16 ms frame rate from the observation sequence. Moreover, in each segment, the kernel matrix is constructed by a moving shift window (length of 5 ms) with kernel function. The parameter setting for the baseline experiment (energy level based VAD with Otsu's method for threshold selection) is the same as used in [9]. Before doing the VAD experiments for detection rate evaluation, we show some examples to see the effect of the discriminative ability

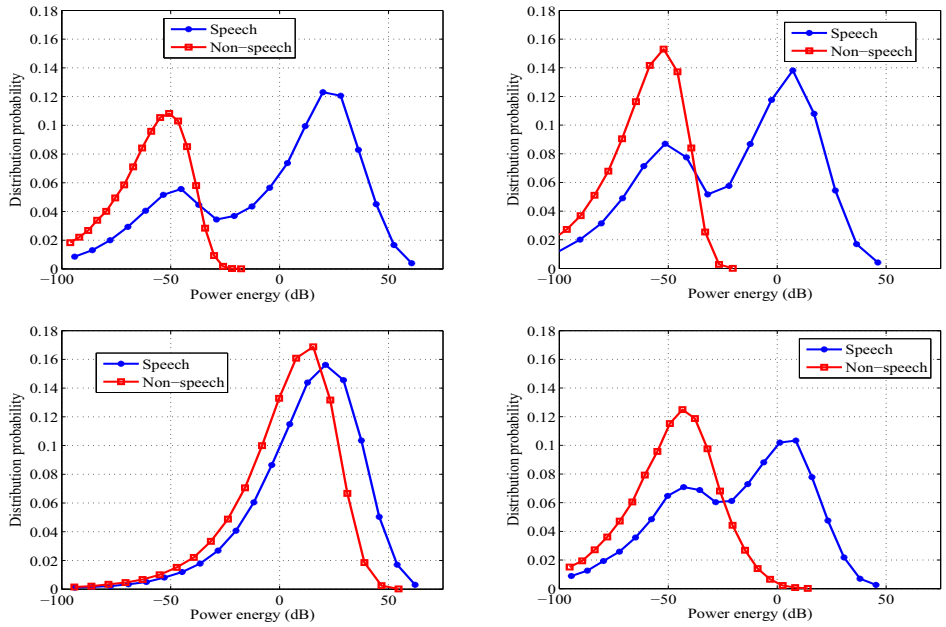

Figure 1: Probability distributions of the log power energy of speech and non-speech in the original input space for the clean (upper-left) and noisy (lower-left) utterances, in the regularized RKHS for the clean (upper-right) and noisy (lower-right) utterances.

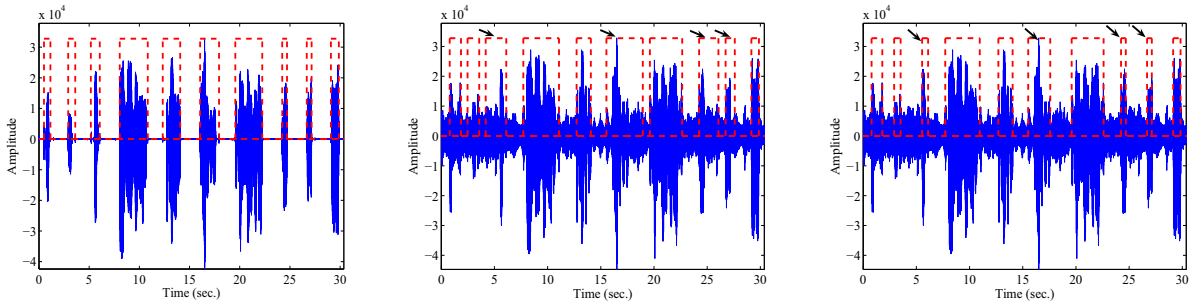

Figure 2: VAD for a clean speech in the original input space (left panel), and noisy utterance (SNR=10 dB) in the original input space (middle panel) and the regularized RKHS (right panel).

between speech and non-speech after the proposed processing intuitively.

4.1. Separability of the distributions between speech and non-speech segments

One clean utterance and its noisy one with signal to noise ratio (SNR) 10 dB (train noise) from CENSREC-1-C [11, 9] (concatenation of utterances) were used for VAD test. The utterance has duration about 30 seconds. The speech and non-speech segments were first collected for the clean and noisy utterances based on the reference VAD, respectively. Based on the collected speech and non-speech segments, their distributions of frame log power energy were estimated (normalized histogram of the log energy distribution) in the original input space and regularized RKHS, respectively. The separability of the distributions between speech and non-speech segments can be used as an index to predict the goodness of the VAD algorithm. The distributions are shown in Fig. 1 for the original input space (left column) and regularized RKHS (right column). Comparing the two panels in the left column of Fig. 1, we can see that for the noisy condition in the original input space there are large overlaps of the probability distributions between the speech and non-speech clusters. Large misclassification will occur (large false alarm for speech and non-speech detections) for the VAD designed in this space. Comparing the two panels in the right column of Fig. 1, we can see that even for noisy condition, the good separability of the distributions between the speech and non-speech clusters is kept well in the regularized RKHS. Intuitively, we can expect a robust VAD performance in this RKHS.

An example of the VAD results for a clean and noisy utterances (SNR=10 dB) in the original input space and the regularized RKHS are shown in Fig. 2. Comparing the VAD result for clean speech (left panel) and noisy speech (middle panel), we can see that in the original input space, several speech segments are not accurately detected in noisy environments. But as shown in left panel of this figure, we can see that the detections of speech segments are more accurate around the marked periods (the VAD results are labeled on the noisy waveform for the convenience of comparison), i.e., in the regularized RKHS, the performance of VAD is better than that of in the original input

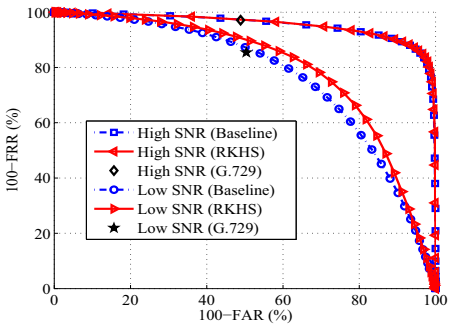

Figure 3: ROC curves of the VAD algorithms for high and low SNR conditions.

space.

4.2. Voice activity detection experiments

In our VAD experiments, the CENSREC-1-C data corpus is used which is a Japanese continuous data corpus (digit strings) designed for testing VAD algorithms in noisy environments [9]. Two data sets, i.e., set A and set B, are used. Set A is composed of four noisy conditions of subway, babble, car and exhibition noise, and set B is composed of another four noisy conditions of restaurant, street, airport, and station noise. In the testing, two types of SNR conditions are used, i.e., high SNR condition which is composed of noise conditions with SNR 20, 15, and 10 dB, and low SNR condition which is composed of noise conditions with SNR 5, 0 and -5 dB. In each SNR condition, there are 104 speech data files. The frame level evaluation measure is used in testing the VAD algorithms. In this evaluation measure, two indexes named as False Rejection Rate (FRR) and False Acceptance Rate (FAR) are defined as follows:

$$\mathrm{FRR} \triangleq \frac{N_{\mathrm{FR}}}{N_{\mathrm{s}}} \times 100 \qquad (\%) \qquad (10)$$

$$\mathrm{FAR} \triangleq \frac{N_{\mathrm{FA}}}{N_{\mathrm{ns}}} \times 100 \qquad (\%) \qquad (11)$$

In Eqs. (10) and (11), the N_{s}, N_{ns}, N_{FR}, and N_{FA} are the total number of speech frames, the total number of nonspeech frames, the number of speech frames detected as non-speech frames, and the number of nonspeech frames detected as speech frames, respectively. By varying the threshold as defined using Otsu's method [9], we calculate the VAD results, and measure the performance based on the FRR and FAR. We average all the results for all noise types for the high and low SNR conditions. The final performance evaluation is represented as the receiver operating characteristic (ROC) curve. For comparison, the VAD in the original input space based on the Otsu's method, and G.729B VAD method [10] are also used. The results are shown in Fig. 3. In this figure, the x-axis is the value of 100-FAR, and the y-axis is the vaule of 100-FRR. From this figure, we can see that in high SNR condition, the performance is almost similar for the baseline VAD and regularized RKHS based VAD, as well as the G.729B VAD (only one diamond-point in the 100-FAR and 100-FRR coordinate). In the low SNR condition, all the performances degrade compared with those in the high SNR condition. The G.729B VAD (the star-point) performs a little lower than that of the baseline VAD. Our proposed VAD in the regularized RKHS, performs the best among the compared three algorithms.

5. Conclusion and discussions

In this study, we proposed an RKHS based method for VAD. In the RKHS, we regarded the estimation of clean speech from noisy observations as a functional approximation problem, and by introducing the smoothness constraint of the mapping function in the RKHS, we could obtain a mapped space in which most of the speech information is kept while noise information is smoothed. Based on the algorithm in the RKHS, we did not need the mapping function explicitly by only introducing a kernel function constructed from the observation signal. By choosing the kernel function, we could easily incorporate the nonlinear and high-order statistic information of the signal in the features. Our preliminary experiments showed that the proposed VAD algorithm could outperform the traditional VAD algorithms.

In the proposed algorithm, several problems need to be further investigated. First of all, the parameter selection problem, for example, the regularization parameter λ in Eq. (4), the kernel function $K(\cdot,\cdot)$ in Eq. (5). In our study, these parameters were manually chosen with reference to the final VAD results. In addition, considering the non-stationarity problem of the noise, we need to find an adaptive algorithm to update the construction of the kernel matrix. In the future, we will further develop our algorithm by considering all these questions.

6. Acknowledgements

This study is supported by the MASTAR project of the Knowledge Creating Communication Research Center of National Institute of Information and Communications Technology (NICT), Japan.

7. References

[1] Ramrez, J., Grriz, J. M., Segura, J. C., "Voice Activity Detection. Fundamentals and Speech Recognition System Robustness", in M. Grimm and K. Kroschel. Robust Speech Recognition and Understanding, 1-22, ISBN 978-3-902613-08-0, 2007.

[2] Sohn, J., Kim, N. S., Sung, W., 1999, "A statistical model-based voice activity detection", IEEE Signal Proc. Lett., 6(1):1-3, 1999.

[3] Kato, H., Ishizuka, K., Fujimoto, M., "A Voice Activity Detection Based on an Adjustable Linear Prediction and GARCH Models", Speech Communication, 50(6), 476-486, 2008.

[4] Ishizuka, K., Nakatani, T., Fujimoto, M., Miyazaki, N., "Noise Robust Voice Activity Detection Based on Periodic to Aperiodic Component Ratio", Speech Communication, 52(1), 41-60, 2010.

[5] Fujimoto, M., Ishizuka, K., "Noise Robust Voice Activity Detection Based on Switching Kalman Filter", IEICE Transactions on Information and Systems, E91-D(3), 467-477, 2008.

[6] Kimeldorf, G., Wahba, G., "Some results on Tchebycheffian Spline Functions", J. Mathematical Analysis and Applications, 33(1):82-95, 1971.

[7] Scholkopf, B., Smola, A. J., Learning with Kernels, the MIT Press, Cambridge, MA, USA, 2002.

[8] Otsu, N., "Threshold selection method from gray-level histograms", IEEE Trans. Sys. Man. Cyber., 9:62-66, 1979.

[9] Kitaoka, N., et al, "CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments", Acoustic. Sci. & Tech., 30(5):363-371, 2009.

[10] ITU-T, A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70, Recommendation G.729 Annex B, 1996.

[11] http://sp.shinshu-u.ac.jp/CENSREC/en/CENSREC/CENSREC-1-C/