The Robustness Study of Multiple Kernel Learning Approaches for VAD

Jie Zhang [1, 2], Mantao Wang [1, 2, a*], Haitao Tang [3, b], Qiang Huang [1, 2], Haibo Pu [1, 2], Lixin Luo [1] and Zhihao Zhou [1]

[1] Sichuan Agricultural University, College of Information Engineering, 625000 Yaan, China

[2] The Lab of Agricultural Information Engineering, Sichuan Key Laboratory, 625000 Yaan, China

[3] Harbin Institute of Technology, Speech Processing Lab, 150001 Harbin, China

[a]wangmantao@sicau.edu.cn, [b]HaitaoTang.Ape@gmail.com

Keywords: Voice activity detection; Deep learning; Multiple Kernel Learning; Robustness

Abstract. Recently, although the MKL-SVM-based VAD has achieved desirable performance, the VAD base on deep learning networks, are attracting greater research interest than their with overwhelming advantages. In this paper, we focus on investigation and analysis the noise robustness of VAD systems multiple-feature-based on MKL-SVM comparing DBN, LSTM and CNN at frame level under various noisy conditions on TIMIT. Experimental results have shown that the MKL-SVM-based VAD not only is not inferior to deep learning networks VADs, but also has a low detection complexity. Further experiment on the information robustness task demonstrates that the MKL-SVM-based VAD apply the advantages of multiple features effectively.

Introduction

The voice activity detection (VAD), which refers to the detection technique of automatically distinguishing active speech from non-speech regions, has become one of the significant components in speech signal processing and has been applied in many applications. In recent years, with the increasing demands of speech processing raised from various practical services, the VAD technique has to face new challenges in the low signal to noise ratio (SNR) and non-stationary noise environments. Although numerous methods with great achievements have been developed in the VADs, there still exists a large gap between the present results and actual system requirements.

In a clean signal, or one that has high SNR, the VAD problem can be solved directly using methods mentioned above. However, when the signal is corrupted by noise, it is difficult to distinguish between speech and non-speech. Enqing et al. [1] concatenated the acoustic features used in the G.729B VAD [2] in serial, and proposed to apply single kernel support vector machine (SVM) to VAD for the first time in 2002, where the kernel-induced feature mapping is further utilized to enhance the discriminability of the learning machine. It achieves significant improvement over the G.729B VAD. Then SVM-based VAD has been applied widely by many researchers. However, more recently deep learning approaches are attracting greater research interest than others with their overwhelming advantages in VAD. For example, Paper [3] presented that deep belief networks (DBN)-based VAD system can fuse the advantages of multiple features much better than traditional VADs. Eyben and Weninger [4] proposed a VAD approach based on recurrent neural networks (RNN) and long short-term memory (LSTM), which takes advantage of its ability to model long range dependencies between the inputs and improve the robustness in real-life applications. Besides, convolutional neural network (CNN) is also applied to VAD by Thomas et al. [5], since CNN can achieve stronger feature vectors that are more invariant to input distortion and position and is easier to train due to the parameter sharing [6]. Although most single kernel SVM-based VADs do some efforts, the main advantage of these VADs is still lying in the superiority of the SVM-based approaches to the deep-learning-based approaches during this time.

In this paper, we focus on a kind of multiple kernel learning support vector machine (MKL-SVM), and further extended it to the supervised MKL-SVM for the multiple-feature-based VAD. Although the multiple kernel support vector machine has achieved desirable performance in [7], the lack of thorough comparisons and analysis with those deep learning methods makes people still unaware of the advantages and disadvantages on VAD task, especially under unseen noisy environment and low SNR conditions. In this paper, we investigate and analyse the noise robustness of VAD systems based on MKL-SVM comparing DBN, LSTM and CNN at frame level under various noisy conditions on TIMIT, a commonly used acoustic-phonetic continuous speech corpus.

The Multiple Kernel Learning Support Vector Machine for VAD

So far the MKL is mostly applied by linear combination. Although the kind of combination may cause the loss of speech information, it has the huge time and space complexity for multiple kernels doing non-linear combination. Therefore most of MKL approaches commonly apply multiple kernel linear combination, because it has smaller cost comparing to develop the new kernel functions.

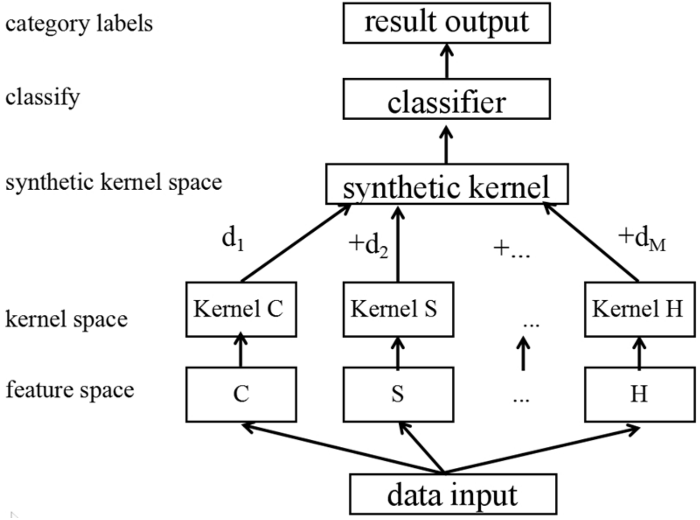

Figure 1: Process schematic of multiple functions of linear combination.

Fig. 1 shows that the process schematic of multiple functions of linear combination. There are many multiple kernel linear combination approaches, such as direct summation kernel, weighted summation kernel, weighted polynomial extension kernel. A kind of kernel function can get different performance combining various speech features. And the selection of kernel functions is decided by the demand of speech processing. So various combines of different speech features can obtain different performance. However, the issue of selection on the kernel functions is one of the most important keys. More recently, the most common and the most widely used is weighted summation kernel, we get the following

$$G=\sum_{m=1}^{M} d_m K_m(x,x'), d_m \ge 0, \sum_{m=1}^{M} d_m = 1 \qquad (1)$$

where x is the support vectors of speech features that the multi-dimension can be shown in Table 1. Each sample which represents speech frame is obviously continuous in the x. The support vectors are obtained from the training sample through an optimization process, and therefore they are a subset of the training sample. And frame-based classification is performed by a comparison among posterior probabilities of the two classes for each frame. G is the kernel function which is in the synthetic kernel space, and $K_m(x,x')$ is the kernel function which performs implicit mapping into a high-dimensional feature space. Gaussian kernel and polynomial kernel primarily used to be the kernel functions in most researches. M is the number to synthesize final kernel. d_m is defined the weight coefficient of kernel matrix.

We apply the Simple MKL to obtain weight coefficient. According to the above functional framework and inspired by the multiple smoothing splines framework of Wahba, we propose to address the MKL-SVM problem by solving the following convex problem, which we will be referred to as the primal MKL problem:

$$
\begin{aligned}
&\min_{w,b,\xi} \frac{1}{2}\sum_{m=1}^{M}\frac{1}{d_m}\left\|w_m\right\|^2 + C\sum_i \xi_i \\
&s.t\ \ y_i\Big(\sum_{m=1}^{M}(w_m \cdot \phi_i(x)) + b\Big) \geq 1-\xi_i \\
&\xi_i \geq 0, \forall i \\
&\sum_{m=1}^{M} d_m = 1, d_m \geq 0, \forall m
\end{aligned} \tag{2}
$$

where each d_m controls the squared norm of w_m in the objective function. The smaller d_m is, the smoother w_m (as measured by $\left\|w_m\right\|$) should be. When $d_m = 0$, $\left\|w_m\right\|$ has also to be equal to zero to yield a finite objective value. The ℓ_1-norm constrain on the vector d is a sparsity constraint that will force some d_m to be zero, thus encouraging sparse basis kernel expansions.

Experiments

In this important section, we will compare the robustness and efficiency of the proposed MKL-based VAD with DBN, LSTM, CNN which referenced VADs, and we will also focus on robustness ability of the MKL-based VAD with respect to different SNRs and various types of noises.

Dataset. Four noisy test corpora of NOISEX 92 0 which is added to the clean speech signals with different SNRs is used for performance analysis. Three different SNR levels of the speech signals are selected, which are [-5, 0, 5] dB respectively. Therefore, there are totally 12 test corpora used for analysis and evaluation. Each test corpus of TIMIT 0 contains 1000 utterances, which is regularly used for evaluation of speech recognition systems and contains only clean speech data. TIMIT contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. Those utterances are split randomly into two groups for training and test respectively. Each training set consist of 700 utterances and each test set consists of 300 utterances. Note that the corpora in the same background noise scenario but at different SNR levels are split with the same random seed, and have the same manual labels. In this paper, the sampling rate is 16 kHz. We set the frame length to 20 ms long, which means that each frame consists of 200 samples.

We concatenate all short utterances in each acoustic data set to a long one so as to simulate the real-world application environment of VAD. Eventually, the length of each long utterance is in a range of (180, 240) seconds long. Because speech can be approximated as a stationary process in short-time scales, we usually divide speech signals into a sequence of overlapped short-time frames with all frames set to an equivalent length of 10 to 20 ms long. The frame is used as the basic detection unit in most cases, such as SVM-based VAD. However, deep learning-based VADs compare the effectiveness and efficiency with the single kernel SVM achieving significant improvement. Hence the VAD problem can be partly viewed as a comparative study of robustness of MKL-SVM and deep learning networks. Note that the real-world working environments of VAD are rather complicated, the MKL-SVM-based approach is a challenging topic.

Acoustic Features for VAD. To better show the advantages of the nine different features, we extract every acoustic feature from each observation. They are short-term energy, zero crossing rate, maximum autocorrelation peak, main-vice peak ratio, fundamental frequency, mel frequency cepstrum coefficient, cepstrum coefficient, pectrum variance and spectral entropy. All features are normalized into the range of in dimension 0.

Parameter Settings. We compare the MKL-SVM-based VAD with 3 VADs, which cover a broad deep learning networks-based research area of VAD.

In order to directly compare the performances of DBN, LSTM, CNN the amount of parameters was fixed to almost the same. In all following experiments, the input layers were formed from a context window of 11 frames creating an input layer of 60 visible units. The hidden layer structure of DBN was 437[units]× 4[layers] while it was 150[LSTM blocks]× 1[layer] for LSTM. For CNN, the entire network layer is 2[24×11 input maps], 64[7×2-filter and 2×2-pooling maps], and 3 fully-connected layers of 64, 54, 150 units, respectively. A final layer of 2 units included speech and silence class. By such configuration, the numbers of parameters of DBN, LSTM and CNN are 1345, 1358, and 1369, respectively. That means the parameters of LSTM is nearly equal to CNN's and DBN's. However, for the proposed MKL-SVM-based on Simple MKL, we regard the serial combination of all 9 features as a new feature. Minimum classification error (MCE) is used as the objective of the structural MKL-SVM. The regularization parameter ℓ is searched from $\{2^9, 2^{10}, ..., 2^{14}\}$. Each acoustic feature x_q uses different kernels, with kernel widths being $\{2^{-1}\gamma_q, 2^0\gamma_q, 2^1\gamma_q\}$, respectively, where γ_q is the average Euclidean distance between the observations of the acoustic feature. Therefore, in order to directly compare the robustness of the kernels with different types and quantity, we conduct numerous comparative experiments in the Table 1.

Table 1 Comparison of the robustness of the kernels with different types and quantity in clear speech. SNR in all combination is above 5 dB.

Kernel			Accuracy (%)
Linear	Polynomial	RBF	
15	0	0	84.21
25	0	0	86.04
0	15	0	86.73
0	25	0	88.35
0	0	15	88.85
0	0	25	90.43
5	5	5	93.07
10	10	10	96.46

The results are listed in Table 1. From the table, we can see that the RBF which is also named Gaussian kernel can obtain the better preference comparing with the linear and polynomial. The following two explanations might be reasonable. The first reason is that the simple of speech features can be ideally mapped to a much higher dimensional space by RBF kernel. Another is that kernel parameters directly affect the complexity of function. And the RBF needs the less parameters than polynomial. We also observe that the less quantity of kernel which show the potential of the more simple structures show a poor generalization ability and performs terrible under 5 SNR in clean speech. Moreover, the mixture combination compare the accuracy of VAD with the single, it has the best performance due to the more complex structures. Therefore, we totally use respectively 10 base linear, polynomial, RBF kernels. Note that in the supervised fine-tuning phase of Simple MKL, we run 100 epoches thoroughly and pick up the model that achieves the highest accuracy on the development set from all 100 models without considering the early stopping scheme.

4) Comparison Schemes: We run all experiments 5 times and report the average performances. We evaluate the significant statistical difference of the performances via the two-tailed t test with a confidence interval at 95%.

Because over 55% frames are speech, we use the detection accuracy of MCE as the evaluation metric due to inexistence of reporting misleading results caused by two kinds of classes imbalance.

Table 2 Accuracy (%) Comparison of the Referenced VADs.

Noise Type	SNR	Simple MKL	DBN	LSTM	CNN
Clean	-5	89.23	88.45	89.25	88.01
	0	93.78	92.45	93.66	91.63
	5	96.46	95.99	96.23	94.81
Babble	-5	87.02	85.45	86.45	84.84
	0	90.79	88.48	89.87	87.38
	5	93.34	91.99	93.32	90.87
Factory	-5	88.02	86.45	88.33	86.43
	0	91.87	89.65	92.46	89.07
	5	95.01	93.19	96.39	91.06
Volvo	-5	87.93	86.56	87.85	86.46
	0	90.85	89.39	91.49	88.56
	5	94.75	92.41	95.21	90.26
White	-5	88.15	87.43	88.89	87.32
	0	92.54	91.43	92.58	89.99
	5	95.89	94.39	95.79	92.32

Table 3 Average CPU Time (in minutes) of the MKL-SVM and Deep Learning Networks-based VADs Over Different SNR Levels

Training Time				
Noise Type	Simple MKL	DBN	LSTM	CNN
Clean	325.45	320.38	339.24	324.56
Babble	326.94	323.47	340.76	325.92
Factory	329.72	324.76	341.52	326.72
Volvo	332.64	326.79	342.69	327.19
White	325.47	322.58	340.04	325.73
Test Time				
Noise Type	Simple MKL	DBN	LSTM	CNN
Clean	13.23	13.56	13.64	13.57
Babble	13.34	13.65	13.72	13.68
Factory	13.36	13.69	13.77	13.69
Volvo	13.54	13.71	13.81	13.71
White	13.33	13.59	13.69	13.62

Comparison of Robustness. In the subsections, we try to show the advantage of the MKL-SVM-based VAD empirically via a broad experimental comparison with other VADs. Table 2 lists the accuracy comparison between the deep learning-based VADs and the MKL-SVM-based VAD. From the table, we observe that the MKL-SVM-based VAD is significantly better than the CNN-based VADs. However, the DBN-based VADs does not also yield a better performance than it, and even suffer from slight performance degradations. Moreover, the LSTM-based VAD outperforms other deep learning-based VADs, which shows the potential of the deeper models. In any case, it obtains the same effectiveness with the proposed MKL-SVM-based VAD. The following two explanations might be reasonable.

One possible explanation is that MKL is good at finding the latent manifold of a highly variant problem, such as handwriting, speech, face and topic recognition in natural language processing, but when the manifold characteristic of the problem is relatively apparent, it will not be much better than a well tuned shallow model. Hence, in the future work, we should pay particular attention to the acoustic features that are developed from physical and physiological areas for the diversity between the features [11, 12].

Another possible explanation is that concatenating all features to a long feature vector and further using the multiple kernels combination in different functions might not be the most effective topological hyperplane structure, since different features might be good at reflecting different local patterns of the time and spacial distributions of speech [13, 14]. Therefore, in the future, we should also concentrate on designing effective multiple kernels combination.

Comparison of Efficiency. In this subsection, we focus on the CPU time comparison between the MKL-SVM and deep learning-based VADs, since they use the same input. The results are listed in Table 3. From the table, we can see that MKL and deep learning networks have comparable training time, while MKL is partly more efficient than deep learning networks in prediction. In other words, some of deep learning networks need the same training time than the Simple MKL. The reasons are as follows:

In respect of deep learning networks, the more layer we use, the longer the training and test time of them will be. In our experiments, 30 kernels which based on linear, polynomial, RBF kernels are used in MKL. Although we do use a large number of kernels for the state-of-the-art performance of MKL-SVM, so many kernels is responsible for the inefficiency of MKL-SVM. However we might lower the time complexity of MKL by changing the topology of kernels without suffering a performance degradation. From this point, the MKL-SVM-based VAD may be superior to the deep learning networks. Anyway, both methods meet the real-time detection demand of VAD under the parameter settings of this paper.

Conclusion

In this paper, VAD systems based on MKL-SVM are thoroughly compared from the robustness aspect with different deep learning approaches, such as DBN, LSTM and CNN. Through a series of experiments on TIMIT, it is demonstrated that MKL-SVM is more robust than DBN and CNN under various circumstances and obtains the same robustness with LSTM. Moreover, another key advantage of this introduction is also that the multiple kernel learning can combine multiple features in a nonlinear way. Experimental results have shown that the MKL-SVM-based VAD not only is not inferior to deep learning networks VADs, but also has a low detection complexity. Further experiment on the information robustness task demonstrates that the MKL-SVM-based VAD can apply the advantages of multiple features effectively. Although MKL-SVM-based approaches of VAD performed well under noise-matched conditions, very large performance degradations were observed in conditions with different noise or very low SNR for the proposed approaches. We are looking forward to obtaining the better performance under the kind of conditions in the future research.

Acknowledgements

This work was supported in part by the Youth Fund of the Sichuan Provincial Education Department under Grant 18ZB0467 , and in part by Special Research on Social Sciences of Sichuan Agricultural University in 2017.

References

[1] E. Dong, G. Liu and Y. Zhou: *International Conference on Signal Processing* (September 9-11, 2002), Vol.2, p.1124-1127.

[2] A. Benyassine, E. Shlomot and H.Y. Su: IEEE Communications Magazine, Vol. 35 (1997) No.9, p.64-73.

[3] X. Zhang, J. Wu: IEEE Transactions on Audio Speech & Language Processing, Vol. 21 (2013) No.4, p. 697-710.

[4] F. Eyben, F. Weninger and S. Squartini: *IEEE International Conference on Acoustics, Speech and Signal Processing* (November 15-18, 2003), Vol.32, p.483-487.

[5] S. Thomas, S. Ganapathy: *IEEE International Conference on Acoustics, Speech and Signal Processing* (March 19-22, 2014), p.2519-2523.

[6] Y. Bengio, Y. Lecun: Handbook of Brain Theory & Neural Networks, Vol. 8 (1995) No.12, p.145-148.

[7] J. Wu, X.L. Zhang: Signal Processing Letters IEEE, Vol. 18 (2011) No.8, p.466-469.

[8] J.S. Garofolo, L.F. Lamel and W.M. Fisher: Nasa Sti/recon Technical Report N, Vol. 6 (1993) No.93, p.190-197.

[9] A. Varga, H.J.M. Steeneken and M. Tomlinson: Technical Report Speech Research Unit Defense Research Agency, Vol. 5 (1992) No.5, p.88-97.

[10] B.C. Kuo, K.Y. Chang: Springer Berlin Heidelberg, Vol. 9 (2005) No.8, p.90-98.

[11] A.S. Bregman: MIT Press, Vol. 11 (1999) No.7, p.101-109.

[12] D.L. Wang, G.J. Brown: IEEE Transactions on Neural Networks, Vol. 19 (2008) No.1, p.199-205.

[13] D.L. Wang, D. Terman: IEEE Transactions on Neural Networks, Vol. 6 (1995) No.1, p.283-289.

[14] K. Hu, D.L. Wang: IEEE Transactions on Audio Speech & Language Processing, Vol. 21 (2013) No.2013, p.122-131.