Hidden Markov Models in voice activity detection

Jiří Tatarinov, Petr Pollák

Czech Technical University, Faculty of Electrical Engineering
CTU FEE K13131, Technická 2, 166 27 Prague, Czech Republic
jiri.tatarinov@atlas.cz, pollak@feld.cvut.cz

Abstract

This paper describes two algorithms for speech/pause detection based on Hidden Markov Models. There are proposed algorithm based on separate training and testing and algorithm on simultaneous training-testing procedure. Algorithm are compared to the differential cepstral detector. HMM based algorithms are successful especially under low SNR . Under higher SNR they reach comparable results to the cepstral detector.

1. Introduction

Many systems for noisy speech processing usually require reliable speech/pause detector. While energy based detectors often fail, cepstral ones give better results. Because we the cepstral detectors are commonly used, they can serve as a good reference point. In speech recognition are often used Hidden Markov Models and recognizers based on them are successful. In this paper, we use HMMs for the related but another task - detection of voice activity.

2. Speech/pause detection

There are two approaches to speech/pause detection using HMM. The first one is based on separated training of HMM and following testing. After training the HMMs will not alter its parameters and the model will be insensitive to changes of environment - detection without adaption. But what will happen if the environmet will change? This problem should solve the second approach, which combines steps of training and testing - detection with adaption.

In both procedures we should decide what type of HMM to use. We choosed continuous HMMs with 3-states without mixtures, but this parameter can be object of further considerations. Important is also signal processing - for each frame we calculated cepstral coefficients. There are used two HMMs - model of silence and model of speech.

This is the simplest version of VAD using HMM. The algorithm consist of following steps:

1. *Initialization*
   Set randomly initial models of silent λ_N and speech λ_S .

2. *Training*
   Models λ_N and λ_S are trained on hand-labeled data using Baum-Welch algorithm. Thus model $\lambda = (A, B, \pi)$ is reestimated in this way:

$$\bar{\pi}_j = \gamma_1(i) \tag{1}$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi(i,j)}{\sum_{t=1}^{T-1} \gamma(i)} \tag{2}$$

$$\bar{b}_j(k) = \frac{\sum_{\substack{t=1 \\ o_t = v_k}}^{T-1} \gamma(j)}{\sum_{t=1}^{T-1} \gamma(j)}. \tag{3}$$

   Where $\gamma_t(i)$ is the probability of being in state i at time t and $\xi_t(i,j)$ is the probability of being in state i at time t . In depth is this algorithm described in [5].

3. *Criteria function*
   After training on training set we switch to the testing set. For each frame n we compute $log(P(\mathbf{O}|\lambda_S))$ and $log(P(\mathbf{O}|\lambda_N))$ using forward procedure. As a criteria function $c[n]$ we consider function defined as

$$c[n] = log(P(\mathbf{O}|\lambda_S)) - log(P(\mathbf{O}|\lambda_N)). \tag{4}$$

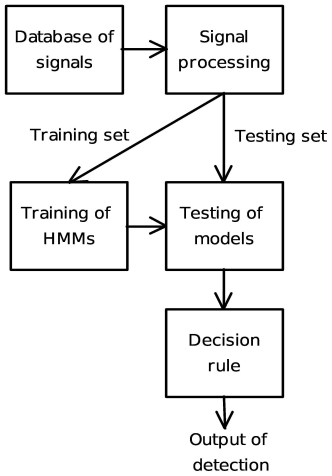

Figure 1: Scheme of HMM voice activity detector without adaption

4. *Calculating of the threshold*
   There are selected a % of the lowest values and b % of the highest values. From these vectors we calculate means μ_a and μ_b . These values determine the dynamic range. Threshold Thr is then calculated as

$$Thr = l(\mu_b - \mu_a) + \mu_a \tag{5}$$

   where l should be in range from 0 to 1 .

5. *Detection*
   Segments where $c[n] > Thr$ are labeled as speech an segment with $c[n] < Thr$ are labeled as silence.

The diagram of HMM voice activity detector is on figure [1].

2.1. Detection with adaption

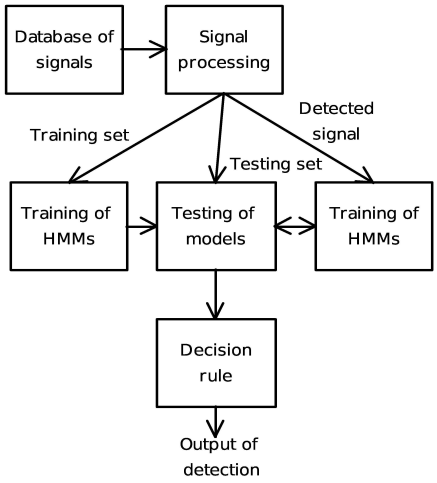

Figure 2: Scheme of HMM voice activity detector with adaption

1. *Initialization*
   Set randomly initial models of silent λ_N and speech λ_S .
2. *Training*
   Models λ_N and λ_S are trained on hand-labeled data using Baum-Welch algorithm.
3. *Updating of the silent model* λ_N
   On the testing set of signals we suppose there is the silent on the beginning of each signal. We update model λ_N using these initial frames.
4. *Criteria function* for adaption
   For each frame n we compute $log(P(\mathbf{O}|\lambda_S))$ and $log(P(\mathbf{O}|\lambda_N))$ using forward procedure. As a criteria function $c[n]$ we consider function defined as

$$c[n] = log(P(\mathbf{O}|\lambda_S)) - log(P(\mathbf{O}|\lambda_N)). \tag{6}$$

5. *Calculating of the threshold*
   There are selected a % of the lowest values and b % of the highest values. From these vectors we calculate means μ_a and μ_b . These values determine the dynamic range. Threshold Thr is then calculated as

$$Thr[n] = l(\mu_b - \mu_a) + \mu_a \tag{7}$$

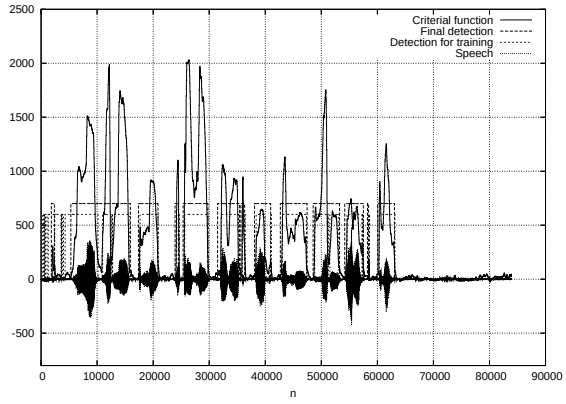

Figure 3: Illustration of speech/non-speech detection using algorithm with adaption

   where l should be in range from 0 to 1 . This threshold is calculated continuously as progress the detection. In the beginning is very small and as speech starts it increases its value and further is not varying too much.

6. *Updating of the model* λ_n
   If the $c[n] > Thr[n]$ there will be not adaption of λ_n , if $c[n] < Thr[n]$ there will be the adaption of λ_n using Baum-Welch algorithm.
7. *Criteria function*
   For adaption of the model of silence it should not happen they are updated using the speech segments, so the threshold is adjusted to this condition. For the detection we can use another threshold $Thr2$ computed from the full length of signall.
8. *Detection*
   Segments where $c[n] > Thr2$ are labeled as speech an segment with $c[n] < Thr2$ are labeled as silence. Results of detection are stored in function $d[n]$.
9. *Postprocessing*
   Because the output of detected speech usually contains bad decisions because of fluctuations of beckground noise characteristics we can smooth it using median filtration, i. e.

$$d_m[n] = med(d[n], m). \tag{8}$$

   The order of the median filter m has influence upon detector results. False decisions are removed better for higher values of m but, on the other hand, bad determinations beginnings and ends of speech sequences appear.

Whole process is figured out on figure [3]. There is shown function indicating intervals for training of model λ_n and function indicating detected speech. There is also smoothed criteria function. The scheme diagram is shown on figure [2].

3. Experiments

All algorithms were experimentaly tested on the PC. All experiments were realized with the real signals collected in the telephone database SpeechDat(E) of Czech language. Testing data consist of 100 sentences read by adult men and women. Sentences consists of ten digits, each digit being different. Sections of speech are manually marked.

3.1. Signal processing

All algorithms were developed under constrained that the length of segments has to be 64 samples with 50% overlapping and sampling frequency 8000 Hz. There was used Hann window for segmentation and from each segment we used 5 cepstral coefficient calculated using DFT.

3.2. Classification

We tried to determine reliable objective criteria for the comparison of different algorithms. These criteria are based on the computation of correct detection rates. Particular criteria were established: *correct speech detection rate* - $P(A/S)$ and *correct non-speech detection rate* - $P(A/N)$. Another possibility is the using of global criteria: *correct detection rate* - $P(A)$ and *speech/non-speech resolution factor* - $P(B)$ defined as

$$P(A) = P(A/S)P(S) + P(A/N)P(N) \tag{9}$$

and

$$P(B) = P(A/S)P(A/N) \tag{10}$$

where $P(S)$ and $P(N)$ are rates of speech and pauses in the processed signal.

Detectors were tested under different noisy condition - $SNR = 5 - 30\,dB$. Speech signals were recorded in a silent environment. Noisy speech $x[n]$ we get from original speech $s[n]$ mixed with real car noise $n[n]$ in appropriate rate, i.e.

$$x[n] = s[n] + k \cdot n[n], \tag{11}$$

where k is determined from

$$k = \sqrt{\frac{\sigma_s^2}{\sigma_n^2} \cdot 10^{-\frac{SNR}{10}}} \tag{12}$$

where σ_s^2 and σ_n^2 are powers of signals $s[n]$ and $n[n]$.

3.3. Detector parameters

Successfulness of detectors depend on setting of some constants. We tried adjust values of constants to maximize correct detection rate, which is more objective criterion that correct speech (non-speech) detection rate. For the detector without adaption we setted constants $A = 0.15$, $B = 0.15$ and $l = 0.2$. For the detector with adaption we setted constants for calculating Thr as $A = 0.15$, $B = 0.15$, $l = 0.15$ and for calculating $Thr2$ as $A = 0.15$, $B = 0.15$, $l = 0.05$. Referential cepstral detector used same type of threshold computation with $A = 0.15$, $B = 0.05$, $l = 0.25$. Finaly we used median filter with order $m = 11$.

3.4. Test results

We ran tests using speech without car noise with results shown in table [1]. There are also results of tests on speech with noise shown on figures [4] [5] [6] and [7].

As shown in the table [1], all detectors have comparable results. HMM detector without adaption is better in detection of non-speech segment than in speech segments.

If we look to the results of tests on signal with car noise we see that the best correct speech detection rate reached HMM detector with adaption, the lower SNR is the better results compared to the cepstral detector we got. Correct non-speech detection rate reached both HMM detectors better then the cepstral detector under low SNR . Under higher SNR the HMM detector with adaption had lower correct non-speech detection rate than the others. If we look to the correct detection rate, the best results reached HMM detector without adaption under low SNR . Under higher SNR had this detector comparable result to the referential one. HMM detector without adaption had lower correct detection rate than the others. The highest Speech/non-speech resolution factor reached both HMM detector mainly under lower SNR . Under higher SNR all of compared detectors were comparable.

4. Conclusions

New approach to the voice activity detection comming out of the methods commonly used in speech detection was proposed in this paper. Algorithms are based on Hidden Markov Models. There were proposed algorithm based on separate training and testing and algorithm on simultaneous training-testing procedure. Algorithm were compared to the differential cepstral detector. HMM based algorithms were successful especially under low SNR . Under higher SNR they reached comparable results to the cepstral detector.

5. Acknowledgements

This work supported by GAČR Modelling of Biological and Speech Signalls grant No. 102/03H085.

6. References

[1] Pollák, Petr, Sovka, Pavel and Uhlíř, Jan, “Cepstral Speech/Pause Detectors”, Proceedings of IEEE Workshop on Nonlinear Signal and Image Processing, Neos Marmaras, Greece, June 1995.

[2] Sovka, Pavel and Pollák, Petr, “The Study of Speech/Pause detectors for Speech Enhancement Methods”, Eurospeech, Madrid, 1995.

[3] Zhang, Jianping, Ward, Wayne and Pellom, Bryan, “Phone based voice activity detection using online bayesian adaptation with conjugate normal distributions”.

[4] Jelínek, Tomáš “Speech/pause detector based on a cepstrum derivation”, CTU, 2004

[5] Rabiner, Lawrence R., Juang, B. H. “Fundamentals of speech recognition”, Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[6] Sovka, Pavel and Pollák, Petr, “Vybrané metody číslicového zpracování signálů’, CTU, Prague, 2003.

[7] Murphy, Kevin, “Hidden Markov Model (HMM) Toolbox”, February 2004.

Detector	$P(A/S)$	$P(A/N)$	$P(A)$	$P(B)$
Without adaption	0.7442	0.9170	0.8356	0.6824
With adaption	0.8428	0.7968	0.8185	0.6716
Differential cepstral detector	0.8495	0.8641	0.8573	0.7341

Table 1: Results of experiments on speech without car noise

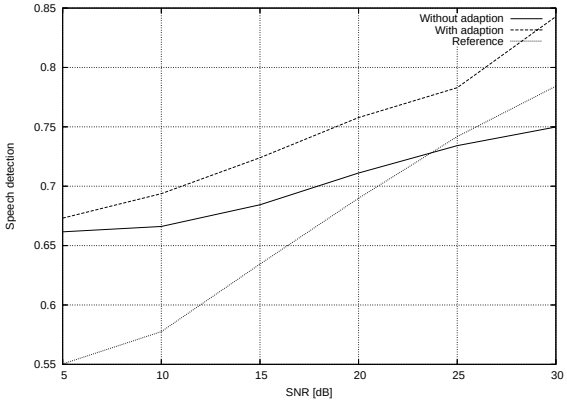

Figure 4: Correct speech detection rate

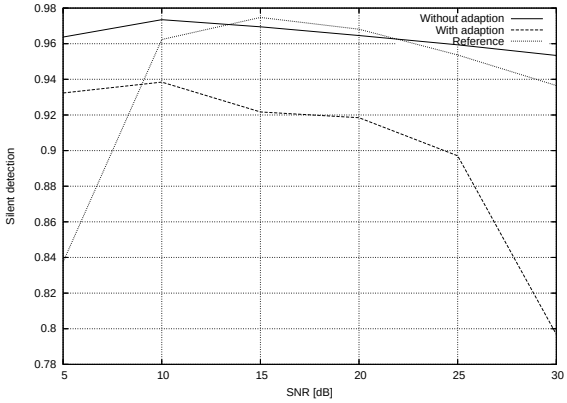

Figure 5: Correct nonspeech detection rate

beginequation20pt]

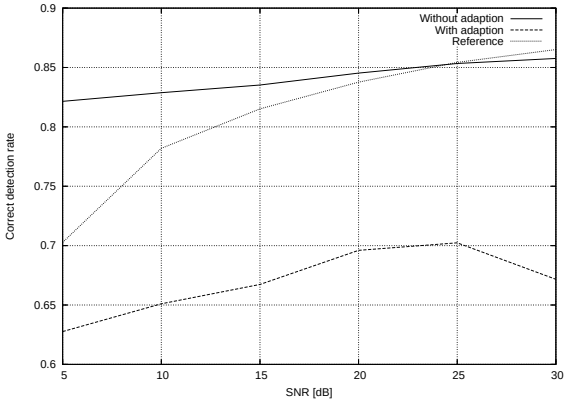

Figure 6: Correct detection rate

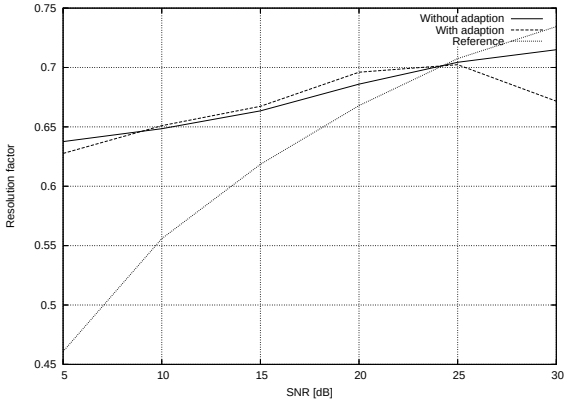

Figure 7: Speech/nonspeech resolution factor