Evaluation and Optimization of Noise Robust Front-End Technologies for the Automatic Recognition of Hungarian Telephone Speech

Péter Mihajlik, Zoltán Tobler, Zoltán Tüske and Géza Gordos

Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Hungary
mihajlik@tmit.bme.hu

Abstract

In this paper a variety of front-end configurations are evaluated on Hungarian telephone speech databases. Our aim was to measure directly the efficiency of the front-ends on real noisy and normal speech data. As a baseline the ETSI ADSR standard front-end is used. Some simplification on the standard is introduced resulting in better performance on our databases than the original front-end in terms of both speed and recognition rate. Besides, another recently proposed feature extraction approach is also investigated. Finally the effect of the novel voice activity detection approach is evaluated. The best front-end configuration augmented with this voice activity detector outperformed significantly the baseline in each recognition test and by 24,7% relative in average.

1. Introduction

Recently many noise robust front-end technologies have been introduced. Complex test databases and methods have been developed in order to give a reliable comparison of the various front-end configurations. However, there is no guarantee that a front-end method proven as the most effective in an Aurora test, will be better than a concurrent approach on independent test data – the results may vary from database to database. Though noise robustness seems to be a language independent issue we decided to evaluate a variety of front-end configurations on Hungarian telephone speech databases. Our goal was to measure directly the efficiency of the front-ends on real noisy speech data therefore no "artificial" noise addition was performed.

We describe our baseline experiments with the ETSI ADSR (Advanced Distributed Speech Recognition) standard front-end [1]. Then some simplification on the standard is introduced which yielded better results on our databases than the original front-end in terms of both speed and recognition rate. Besides, another recently proposed feature extraction approach is also investigated, namely application of the PMVDR (Perceptual Minimum Variance Distortionless Response) [2] coefficients. Finally the effect of a novel VAD (Voice Activity Detection) approach on the recognition accuracy and speed is evaluated, and then we conclude the results.

2. Databases

2.1. Noisy test database

Our general aim was increase the accuracy of our cellular phone-accessible automatic speech recognition systems [3] in the presence of natural and possibly high-energy noises. We found that these noises in such environments as well as the behaviors of the speakers are complex (unexpected noises, Lombard-effect, etc.), and cannot be fully modeled by post noise addition. Therefore we recorded a noisy Hungarian database containing phone calls from typical noisy environments (in-car, shopping center, bureau etc.) [4]. The test set consisted of 1226 spontaneous noisy utterances typically with syllable length of a 2 to 5. A limitation of this database is that some of the recordings are AGC (Automatic Gain Control) distorted.

In one hand the database is real-life so its results can be trusted, on the other hand the SNR (Signal to Noise Ratio) is not controlled. In the followings we will refer to this database to as Noisy DB.

2.2. Normal test database

By the completion of the Noisy DB we were able to test the recognition system on real noisy data. Besides another database containing normal ("clean") telephone speech was prepared for testing in order to check the recognition accuracy in normal conditions. This database was used originally to tune the triphone clustering of our Hungarian ASR system [3]; therefore the phone calls marked as noisy at the annotation process were removed. The test set consists of 6057 utterances – typically read and spontaneous command words. This database will be called Normal DB in the rest of the article.

2.3. Training database

The training database contains normal telephony speech (fixed and mobile, both) with about a length of 3 hours [5]. The noisy recordings were removed again therefore testing on the Normal DB can be referred to as "well-matched" condition and testing on the Noisy DB as "highly mismatched" condition. However, we are going to use the simple "normal" and "noisy" terminology in the followings, because apart from the noises the databases are matched. (Though in the Noisy DB there are numerous AGC distorted recordings absolute energy suppression should compensate this effect)

3. Evaluation methodology

Since usual Aurora-like various SNR and noise type tests were not feasible with our DBs, we applied a simple direct test methodology as follows. Every front-end configuration was evaluated by double running of recognition tests on the Normal and the Noisy DB's. First time standard features were used on both databases (39 dim. of features), second time the absolute energy features like logE+C0, C0, etc. were removed (38 dim. features). As we compared only front-ends, the

single-pass search engine and all other parameters were unchanged. Middle vocabulary command word recognition was performed on both test databases. The vocabulary size was 231 at the Noisy DB and 911 at the Normal DB. No language model was used. The acoustic models were cross-word triphone models trained always on the same way, on the same database.

The bases of the comparison between front-ends were the average relative improvements compared to the baseline. Two values were calculated, one for the standard features and one for the absolute energy suppressed parameters. A configuration was considered better than the baseline only if both average relative improvements were positive. With regard to the size of the test sets two digits relative improvements (in %) was judged as significant.

4. Experiments with various front-end configurations

4.1. Baseline experiments with the ETSI ADSR standard

When evaluating noise robust front-ends it seemed to be a natural choice to use the ADSR standard [1] as baseline. The ADSR front-end showed significantly better performance on the complex Aurora databases than the simple MFCC (Mel-Frequency Cepstral Coefficients) front-end. The improvements were due to additional processing blocks, such as Wiener-filter based noise suppression (NS) [6], waveform processing (WP), and blind equalization (BEQ) [7].

There are also two voice activity detectors integrated into the ADSR front-end. One is used for estimating the noise in the NS block (VADNest) and the other for dropping non-speech frames. The second is based on the first and as our results show it has no significant effect on the recognition accuracy (See Table 1). The processing scheme of the standard is shown in Fig. 1 (solid lines).

The features generated by the front-end are cepstral coefficients (C_0-C_{12}) and logarithmic energy (*logE*) by default. We integrated the server-side feature processing like derivatives and combined energy calculation into the front-end (not shown in Figures) for easier experimentation.

Table 1 shows the recognition results of the ADSR standard, where frame dropping (FD) was disabled and enabled, respectively. The performance of the front-end is unexpectedly poor in the noisy case when the absolute energy was not suppressed. In one hand it may be because of the AGC distortion, but on the other hand we think that the highly variable and intensive noises are big challenges for the ASR system. As frame dropping has no significant effect on the recognition rate it was set as disabled in the following experiments unless the opposite is not mentioned.

Since all data is sampled by a frequency of 8kHz the 16KHz extension of the standard was not evaluated on our databases.

Table 1: Baseline performances (Word Error Rate in %) of the standard front-end

	With energy		Absolute energy suppressed	
	Normal	**Noisy**	**Normal**	**Noisy**
ETSI ADSR	**5,23**	**51,24**	**6,26**	**21,20**
ETSI ADSR + FD	5,21	51,07	6,26	21,20

4.2. Simplifications on the ETSI ADSR standard

The recognition results with the standard front-end seemed somehow not satisfactorily enough therefore we launched a series of experiments with numerous settings of the ADSR front-end. To enable free experimentations each processing blocks was made optional so that e.g. the noise suppression (NS) without waveform processing (WP) could be tested, etc. See Fig. 1 for the block diagram of modified front-end. Among the numerous configurations here we present only the cases when the given configuration cannot be considered as tuning. (So, direct database dependent adjustments like frequency range, etc. were excluded.) To be honest, database specific tuning did not help to outperform the best one of the following results.

First let us see the performance of the pure MFCC (CC) block of the standard in Table 2. A surprisingly high *positive* improvement can be noted in three cases of four. What is unexpected it is the higher relative improvement at the Noisy DB than the Normal DB if the absolute energy was not suppressed. After all this result is not considered better than the baseline ADSR because of the negative average relative improvement in the energy suppression case.

The first configuration that really outperformed the baseline is shown in the next row of Table 2. It is a simple configuration where only the blind equalization (BEQ) block was used after the mel-frequency cepstrum calculation and all other blocks were skipped. All the parameters and adjustments of the original front-end were unchanged.

The best configuration was the one where "half BEQ" was applied. "Half" means using blind equalization only at the test phase and not at the training phase. The results are in the 3rd row of Table 2.

It should be noticed that not only the recognition accuracy was improved but also at the same time the processing time of the feature extraction was decreased. Further details can be found in section 4.6.

A possible explanation of the results may be that the skipped blocks of the ADSR front-end feature like VAD, NS, WP are not well suited for our databases. But since the test data is real telephony speech and language dependency can be excluded we suppose that the ETSI ADSR standard is not generally well applicable for the recognition of noisy telephone speech. However, the relative improvements in case of the best error rates on the Noisy DB were not considered significant therefore we continued with the experimentations.

Table 2: a) Word error rates and b) relative improvements of the simplified configurations compared to the baseline

	With energy			Absolute energy suppressed		
***a)** WER %*	**Normal**	**Noisy**		**Normal**	**Noisy**	
CC	4,78	45,61		5,26	27,33	
CC + BEQ	4,76	43,60		5,43	19,97	
CC + half BEQ	4,38	41,87		4,71	20,63	
***b)** Rel. impr.%*	**Normal**	**Noisy**	**Avg.**	**Normal**	**Noisy**	**Avg.**
CC	+8,6	+11,0	**+9,8**	+16,0	-28,9	**-6,5**
CC + BEQ	+9,0	+14,9	**+11,9**	+13,3	+5,8	**+9,5**
CC + half BEQ	+16,3	+18,3	**+17,3**	+24,8	+2,7	**+13,7**

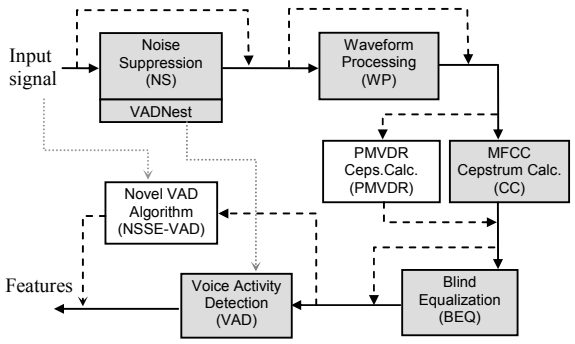

Figure 1: Block diagram of the modified experimental ADSR front-end with optional process flows illustrated by dashed lines (original processing blocks are filled)

4.3. Evaluation of PMVDR feature extraction

[2] proposes the PMVDR approach as a new perspective on feature extraction. It suggests that the PMVDR cepstrum calculation performs better in noisy conditions than MFCC. So, based on the literature we implemented the PMVDR feature extraction method and integrated into the experimental version of the ADSR front-end as a replacement for the CC block (see Fig. 1).

We found that on our databases the optimum parameters of the feature extraction method are $\alpha=0.35$ and $Q=20$.

The recognition results of PMVDR cepstrum calculation in various front-end settings are shown in Table 3. It can be seen that using PMVDR instead of MFCC always resulted in a small degradation of the accuracy. Though the loss of performance seemed to be minor, with regard to the consequently worse results and to the higher computational load of the algorithm (see Table 6) we rejected the PMVDR feature extraction approach as an alternative of MFCC.

Table 3: a) Word error rates and b) relative improvements compared to the baseline front-end

	With energy			**Absolute energy suppressed**		
***a)** WER %*	**Normal**	**Noisy**		**Normal**	**Noisy**	
PMVDR in ADSR	5,49	51,40		6,85	22,28	
PMVDR	4,61	46,08		5,40	27,99	
PMVDR + BEQ	4,75	42,69		5,43	20,21	
PMVDR + half BEQ	4,45	42,44		4,91	20,87	
***b)** Rel. impr.%*	**Normal**	**Noisy**	**Avg.**	**Normal**	**Noisy**	**Avg.**
PMVDR in ADSR	-5,0	-0.3	**-2,6**	-9,4	-5,1	**-7,3**
PMVDR	+11,9	+10,1	**+11,0**	+13,7	-32,0	**-9,1**
PMVDR + BEQ	+9,2	+16,7	**+12,9**	+13,3	+4,7	**+9,0**
PMVDR + half BEQ	+14,9	+17,2	**+16,0**	+21,6	+1,6	**+11,6**

4.4. Evaluation of a novel VAD algorithm

A precise and noise robust endpoint detector can straightforwardly reduce the number of recognition errors by discarding non-speech – but possibly high-energy – sound events of the noisy environment. Therefore we integrated a newly developed high performance Noise Suppressed Spectral Entropy-based Voice Activity Detector (NSSE-VAD) into the experimental front-end.

4.4.1. The NSSE-VAD approach

The method uses spectral entropy instead of energy. In this way voice activity detection can be performed based on simple *global* threshold even in the presence of relatively high-level white noise. However, the standard entropy-based endpoint detection fails if the noise is color and shows some kind of organization.

The essence of our NSSE-VAD algorithm is that the signal spectrum is noise-whitened before the entropy calculation. I.e. the estimated noise spectrum is subtracted form the noisy speech spectrum in the logarithmic domain. This noise suppression step has a particular influence on the entropy curve of the input signal resulting in an outstanding voice activity detection performance [8].

The main steps of the algorithm:

- **Smoothing** the short-time magnitude spectrum of the input signal by convolving with a 2-dimensional smoothing matrix.
- **Estimation of noise spectrum** by searching the minimum values of surrounding frames for each frequency bin, then choosing the maximum between future and past minimums (1-3).

$$S_{past}(bin,t) = \min\left(S(bin,t-B),\ldots,S(bin,t)\right) \quad (1)$$

$$S_{future}(bin,t) = \min\left(S(bin,t),\ldots,S(bin,t+F)\right) \quad (2)$$

$$S_{noise}(bin,t) = \max\left(S_{past}(bin,t), S_{future}(bin,t)\right) \quad (3)$$

  where $B=75$ and $F=25$ covering 1 second of input frames.

- **Noise suppression** or whitening.

$$P_{spec}(bin,t) = \left(\frac{S(bin,t)}{S_{noise}(bin,t)}\right)^2 \quad (4)$$

- **Entropy** calculation

$$H(t) = -\sum_{i=0}^{128} P(i,t)\log\left(P(i,t)\right) \quad (5)$$

  where probabilities *P(i,t)* are calculated from the power spectrum bins as

$$P(i,t) = \frac{P_{spec}(i,t)}{\sum_{j=0}^{128} P_{spec}(j,t)} \quad (6)$$

- **Speech/non-speech decisions** based on a predefined entropy threshold and on simple time constraints (minimum silence and speech segment sizes, hangover and look-ahead, etc.).

The frame dropping is controlled by the output of last stage. More detailed description of the algorithm can be found in [8].

The implementation of the NSSE-VAD algorithm is not optimized yet for speed. Though it is computationally relatively expensive and requires more memory than the standard for example, it is quite effective in dropping non-speech frames so speeds up the recognition procedure.

4.4.2. Recognition results

As it can be clearly seen in Table 4, this novel VAD improved the word recognition rate radically when absolute energy was used, as well. Though not presented explicitly the NSSE-VAD improves the recognition accuracy in every front-end configuration. What is more, the simple MFCC front-end along with the proposed NSSE-VAD performs nearly as well as the ADSR front-end in the highly mismatched, energy-suppressed condition and significantly better in all other test situation.

Finally, it can be seen that the front-end configuration giving the best overall results is the "half BEQ + CC + NSSE-VAD". This means that during training the simple MFCC features were used and at the testing the MFCC front-end was augmented by blind equalization and by the NSSE-based voice activity detection. This setting outperformed significantly the ETSI ADSR standard in each test, and the average improvements were +29,0% and +20,4% when absolute energy was used and not, respectively.

Table 4: a) Word error rates and b) relative improvements of various configurations augmented with NSSE-VAD compared to the baseline (FD i.e., frame dropping was enabled)

	With energy			**Absolute energy suppressed**		
***a)** WER %*	**Normal**	**Noisy**		**Normal**	**Noisy**	
FD, ADSR	5,11	36,14		5,86	20,54	
FD, CC	4,66	35,51		5,08	22,77	
FD, CC+ BEQ	4,70	33,83		5,23	18,65	
FD, CC+ half BEQ	4,27	30,94		4,51	18,48	
***b)** Rel. Impr. %*	**Normal**	**Noisy**	**Avg.**	**Normal**	**Noisy**	**Avg.**
FD, ADSR	+2,3	+29,5	**+15,9**	+6,4	+3,1	**+4,8**
FD, CC	+10,9	+30,7	**+20,8**	+18,8	-7,4	**+5,7**
FD, CC+ BEQ	+10,1	+34,0	**+22,1**	+16,5	+12,0	**+14,2**
FD, CC+ half BEQ	+18,4	+39,6	**+29,0**	+28,0	+12,8	**+20,4**

4.4.3. Frame dropping rates

A comparison was made between frame dropping rates of the ADSR-VAD and the proposed NSSE-VAD. As Table 5 shows the NSSE-VAD dropped an amount of frames more then 2 times larger than ADSR-VAD at the Normal DB. The difference of frame dropping rates of the two VAD is even more convincing at the Noisy database; it is more than a magnitude.

Table 5: Frame dropping rates of the two discussed VAD's on our databases

DB	Detector	Sum of vectors	Dropping rate
Normal	ADSR VAD	1.788.101	24,9 %
	NSSE-VAD		60,0 %
Noisy	ADSR VAD	466.332	3,5 %
	NSSE-VAD		52,6 %

4.5. Processing times of the front-end configurations

In this section we give a very brief overview of processing time measures of the various front-end configurations. The measurements were performed on a 2.2GHz Athlon CPU powered PC.

Table 6: Real-time factors (RTF, % of CPU usage) for the main discussed configurations and their relative value to the ETSI ADSR standards'

Configuration	Real-time factor	Relative processing time
ETSI ADSR	0,0185	100 %
CC	0,0040	22 %
CC + BEQ	0,0040	22 %
PMVDR	0,0077	42 %
NSSE-VAD, CC	0,0240	129 %

It can be seen that the proposed VAD algorithm along with MFCC has a higher computational demand than the ADSR. However, in return the NSSE-VAD has a much higher frame-dropping rate i.e., more than the half of frames was not sent to the recognizer in our experiments. In this way more than 50% of time was saved at the recognition procedure that was much more critical in term of CPU usage at the medium-large vocabulary recognition task than the front-end processing itself.

5. Conclusions

A variety of front-end configurations were evaluated on Hungarian telephone speech databases. Our goal was to measure directly the efficiency of the front-ends on real noisy and normal speech data. As a baseline the ETSI ADSR standard front-end was used. Some simplification on the standard was introduced which yielded better results on our databases than the original front-end in terms of both speed and recognition rate. Besides, another recently proposed feature extraction approach was also investigated, namely application of the PMVDR coefficients. This approach was rejected because standard MFCC-based feature performed better in each test situation. Finally the effect of the novel NSSE-VAD approach was investigated. We found that on our databases the MFCC + half BEQ + NSSE-VAD significantly outperformed the ETSI ADSR in each recognition test by 24,7% relative in average. The computational load of the proposed front-end is moderately higher than that of the ADSR standard, but the extensive frame-dropping rate of NSSE-VAD over compensates this effect.

6. References

[1] *ETSI standard doc.*, "Speech Processing, Transmission and Quality aspects (STQ); Distributed Speech Recognition; … ", *ETSI ES 202 050 v1.1.2.*

[2] *U. H. Yapanel, J. H. L. Hansen*, "A New Perspective on Feature Extraction for Robust In-Vehicle Speech Recognition", *EUROSPEECH, pp. 1281-1284, 2003.*

[3] T. Fegyó et al. "Voxenter – Intelligent Voice Enabled Call Center for Hungarian", *EUROSPEECH, pp. 1905-1908, 2003.*

[4] http://alpha.ttt.bme.hu/speech/hdbtesztelen.php

[5] http://alpha.ttt.bme.hu/speech/hdbMTBA.php

[6] *A. Agarwal, Y. M. Cheng*, "Two-stage Mel-warped Wiener Filter for Robust Speech Recognition", *Proc. IEEE-ASRU workshop 1999.*

[7] *L. Mauuary*, "Blind Equalization in the Cepstral Domain for robust Telephone based Speech Recognition", *Proc. EUSPICO'98, Vol.1, pp. 359-363, 1998.*

[8] *Z. Tüske et al.* "Robust Voice Activity Detection Based on the Entropy of Noise-Suppressed Spectrum" *submitted to InterSpeech 2005*