ON REAL-TIME MEAN-AND-VARIANCE NORMALIZATION OF SPEECH RECOGNITION FEATURES

Pere Pujol, Dušan Macho, and Climent Nadeu

TALP Research Center
Universitat Politècnica de Catalunya, Barcelona, Spain.
[pujol,dusan,climent]@talp.upc.edu

ABSTRACT

This work aims at gaining an insight into the mean and variance normalization technique (MVN), which is commonly used to increase the robustness of speech recognition features. Several versions of MVN are empirically investigated, and the factors affecting their performance are considered. The reported experimental work with real-world speech data (Speecon) particularly focuses on the recursive updating of MVN parameters, paying attention to the involved algorithmical delay. First, we propose a decoupling of the look-ahead factor (which determines the delay) and the initial estimation of mean and variance, and show that the latter is a key factor for the recognition performance. Then, several kinds of initial estimations that make sense in different application environments are tested, and their performance is compared.

1. INTRODUCTION

Feature normalization techniques are able to largely reduce the actual mismatch between training and testing conditions. Both histogram equalization (HE) [1] and mean and variance normalization (MVN) [2] have been proposed to process feature vectors in order to significantly improve the speech recognition performance. MVN is more efficient than HE in terms of both computational load and latency since, by doing the reasonable assumption of Gaussianity, only two parameters are required.

The best recognition results with MVN are usually obtained when the values of mean and variance are previously estimated by averaging along the entire current utterance and are kept constant during the normalization [3]. However, this off-line estimation involves a long delay that is likely unacceptable. Alternatively, when the conditions (environment, channel, speaker, etc.) do not change for a period of time, both mean and variance can be estimated from a given set of previous utterances and so the delay is avoided.

Nevertheless, in certain applications, either the conditions change (like with mobile phones) or previous utterances are not available (like in the beginning of a conversation), so that a real-time processing can only be achieved by using an on-line version of the MVN technique, as those proposed by Viikki and Laurila [2, 4] and used by others [5, 6]. In that approach, mean and variance are recursively updated and a given amount of delay must always be accepted; actually, there exists a tradeoff between delay and recognition error.

In this paper, we report an empirical investigation about several on-line versions of the MVN technique and the factors affecting their performance. Speech recognition experiments are carried out with the Spanish portion of the Speecon [7] database, which includes recordings of each utterance with four different microphone positions, thus allowing us to have three recognition experiments with unmatched conditions like in [8].

This work particularly focuses on the recursive MVN, paying attention to the involved algorithmical delay. First of all, a decoupling of the look-ahead factor (which determines the delay) and the initial estimation of mean and variance is proposed, showing that the latter is a key factor for the recognition performance. Accordingly, several kinds of initial estimations are tested and compared. Also, an improvement of recognition accuracy is shown when, using the information about the first speech-to-non-speech transition, the initial estimates are computed from a mixture of speech and non-speech frames.

2. EXPERIMENTAL SETUP AND BASELINE RESULTS

A subset of the office environment recordings from the Spanish version of the Speecon database [7] is used to carry out the speech recognition experiments with digit strings. The database includes recordings with 4 microphones: a head-mounted close-talk (CT) mic, a Lavalier mic, a directional mic situated at 1 meter from the speaker, and an omni-directional microphone placed at 2-3 meters from the speaker.

125 speakers were chosen for training and 75 for testing, both balanced in terms of sex and dialect. CT recordings are used for training, and we test with all the 4 different microphones, so there are 3 experiments with unmatched conditions.

A SCHMM speech recognition system was used (RAMSES [9]). 76 demiphone models were trained with 2 states per model and 512 gaussians per codebook were used. Three non-speech events - silence, speaker noise, and filled pause - are modelled each by a four-state HMM. When decoding, they are combined in parallel, forming three possible alternatives, each with the same probability.

Once 16 mel-scaled sub-band energies were computed from speech signals sampled at 16 kHz, they were 2nd-order frequency-filtered (FF) [10], and the resulting parameters were used as features, along with their first- and second-order time derivatives. The first-order time derivative of the log energy coefficient was also appended to the feature vector. Results obtained with the described baseline system are presented in Table 1.

	Baseline	Off-line UTT-MVN
CT	98.33	97.94
Lav.	88.21	96.49
1 meter	62.15	94.66
2-3 meters	36.07	76.02

Table 1 Word accuracies (in %) for the four types of microphones.

3. ISSUES INVOLVED IN REAL-TIME MVN

Mean and variance normalization (MVN) is widely used in speech recognition systems to reduce the mismatch between training and testing conditions in presence of additive or convolutional noises [3, 6, 11] Assuming Gaussian distributions, this technique normalizes each component of the feature vector according to the expression:

$$\hat{X}_i[n] = \frac{X_i[n] - \mu_i[n]}{\sigma_i[n] + \theta} \qquad (1.1)$$

where $X_i[n]$ and $\hat{X}_i[n]$ are the *i*-th component of the feature vectors at time frame *n* before and after normalization, respectively, μ_i and σ_i are the mean and variance estimates of the sequence $X_i[n]$ at a given frame, and θ is a floor parameter.

The mean and variance values can be assumed constant along *n* and be estimated by averaging on the entire current utterance. The resulting off-line MVN technique, that is referred as off-line UTT-MVN in this article, is effectively used in speech recognition systems, provided that the involved long delay is acceptable. The second column in Table 1 shows the speech recognition performance yielded by the off-line UTT-MVN in our experiments with Speecon, with θ=0. Note the very large improvement for the three mismatched conditions, though there is a slight loss of performance for the matched (CT) case.

However, when real time is a concern, an on-line MVN algorithm has to be employed [4]. Since the whole utterance is not available in that approach, the time-dependent values of mean and variance used in (1.1) must be computed along the utterance. In the next subsections, two commonly used approaches to determine mean and variance estimates are considered.

3.1. Segment-based updating of mean & variance

In this MVN approach, the mean & variance of the feature vectors are estimated for each *n* within a segment, by using a sliding fixed-length rectangular window centered in the current frame [2]. Therefore, there is a delay of half the length of the window. Recognition results for several window lengths are listed in Table 2. This normalization will be referred as segment-based MVN herewith.

	Sliding window length / Delay (in seconds)			
	4.0/2.0	2.0/1.0	1.0/0.5	0.5/0.25
CT	98.67	97.67	97.66	97.40
Lav.	97.13	96.47	95.20	94.73
1 meter	95.23	93.13	91.14	87.49
2-3 meters	76.63	54.29	47.63	32.80

Table 2 Recognition results using MVN with segment-based updating of mean & variance.

Results show that a delay of 2 seconds (window length of 4s) is needed to obtain results similar to those from off-line UTT-MVN for the furthest microphone. Also, a significant increase of performance is observed for all the microphones when the delay goes from 1s to 2s. As the average length of the beginning non-speech part of utterances is 1.25s, it seems convenient to include a portion of the speech signal in the estimation of mean & variance at frames corresponding to the initial non-speech part**.**

If we accept a delay of 0.25s (last column in Table 2), the segment-based approach can noticeably improve the baseline results for the Lavalier and the 1-meter microphones, but the results with the other two microphones are even worse than the baseline ones.

3.2. Recursive updating of mean & variance

In this approach the mean and variance vectors are initialized using the first *D* frames of the current utterance and then they are recursively updated as new frames arrive. However, for the experiments in this sub-section, a non-causal updating of the mean and variance estimates, which was proposed in [4], is used. In this approach, a *look-ahead* parameter corresponding to the number of frames the frame being used to update the estimations is ahead of the frame being normalized, must be optimized. That look-ahead parameter *D* is the actual delay produced by the normalization technique.

Means and variances of feature $X_i[n]$ are updated at frame *n* by using the recursions

$$\mu_i[n] = \beta * \mu_i[n-1] + (1-\beta) * X_i[n+D] \qquad (1.2)$$

$$\sigma_i^2[n] = \beta * \sigma_i^2[n-1] + (1-\beta) * (X_i[n+D] - \mu_i[n])^2, \qquad (1.3)$$

where β is a forgetting factor. When the end of the utterance is reached by the look-ahead, estimates keep their last value till the last frame of the utterance is normalized.

After preliminary experiments, the forgetting factor and the floor parameter θ were set to 0.992 and 0.001, respectively, to balance the improvements across the various microphones. The non-zero θ value is introduced to avoid numerical problems derived from low standard deviation estimates. Therefore, these values will be used in the forthcoming experiments.

The following experiments aim at study the role of the look-ahead parameter and the initial estimates of mean & variance in the recursive MVN.

3.2.1. Look-ahead in recursive MVN

Our primary objective is to reduce the delay involved in the off-line UTT-MVN or segment-based MVN to a more practical interval while maintaining a high robustness of the technique in noisy conditions. We set *0.25s* as a tolerable algorithmical delay for this purpose. Notice that at this delay the performance of the segment-based MVN is quite low (87.49% / 32.80% for 1m / 2-3m mikes), and this delay is not achievable in UTT-MVN.

Table 3 shows the results using different look-ahead values in the recursive MVN. In this case, the look-ahead parameter coincides with the delay. The entire look-ahead interval is used to calculate the initial estimates of mean and variance; in the case of 0 sec, the first 100 ms are used.

	Look-ahead *D* (in seconds)				
	2.0	1.0	0.5	0.25	0
CT	98.33	98.07	98.13	98.60	97.67
Lav.	97.46	97.27	96.40	96.07	94.94
1 meter	96.41	95.02	92.81	92.72	91.69
2-3 meters	77.21	60.23	51.87	48.77	53.09

Table 3 Comparison of recursive MVN using several look-ahead intervals.

The 0.25s look-ahead case shows considerably better results for all microphones than the same delay in the segment-based MVN (a similar behavior was observed in [4]). However, for achieving performances that are close to the off-line UTT-MVN, larger look-ahead parameters are needed: 0.5s to 1.0s for the first three microphones and somewhere between 1.0 and 2.0 for the 2-3m microphone. Notice a large jump in the 2-3m microphone performance when the look-ahead parameter is larger than 1.25s (the average non-speech interval at the beginning of utterances mentioned before.)

In the way the recursive MVN was implemented in these experiments, the look-ahead parameter controls both the *interval used for the initial estimate* of mean & variance at the beginning of utterance and the *look-ahead interval for the updating* of mean & variance during the application of the normalization throughout the utterance. In order to investigate which of these two factors is more important, in the following experiments we decouple the initial estimation of mean & variance from the mean & variance updating.

3.2.2. Initial estimation of mean & variance in recursive MVN

As a first step in assessing the importance of the initial estimation of mean & variance in recursive MVN, the initial estimates were computed as in UTT-MVN, e.g. from the entire utterance and apart from that, the recursive MVN was applied as in the previous experiments. Interesting enough, we can observe from Table 4 that the performance becomes largely independent on the look-ahead updating interval for all microphones. This reinforces the need of good initial estimates of mean & variance if low delays are desired.

	Look-ahead D (in seconds)				
	2.0	1.0	0.5	0.25	0
CT	98.20	98.20	98.53	98.33	98.34
Lav.	97.27	96.89	96.95	96.68	96.62
1 meter	95.68	95.12	95.19	95.18	95.62
2-3 meters	77.38	77.29	75.72	73.42	75.03

Table 4 Recursive MVN with several look-ahead intervals and UTT-MVN-like initial estimates of mean & variance.

In addition, many scores from Table 4 are better than those from UTT-MVN, which indicates the usefulness of the updating. Notice that if the forgetting factor β in (1.2) and (1.3), equals 1.0, this implementation of the recursive MVN is equivalent to UTT-MVN. In our case β was experimentally set to 0.992, allowing a certain degree of updating of the initial estimates, but still providing smoothed estimates of mean & variance**.**

In the following sections, several strategies of the initial estimation of mean & variance for on-line systems are discussed. They are classified into two categories depending on the data used for computing the initial estimates of mean & variance.

4. INITIAL MEAN & VARIANCE ESTIMATED FROM PAST DATA ONLY

Approaches within this category do not use the current utterance to compute the initial estimates of the mean & variance of the features. Instead, for a given utterance, estimates from the following data sources are used:

- **Current session:** It is assumed that one recording session corresponds to only one speaker and one environment. This condition applies in recognition systems with fixed microphones, environment and speaker. Utterances not included in the test set which were recorded in the current session with a given microphone condition (type and distance from the speaker), are used for computing the initial estimates needed in the parameterization of the test utterances corresponding to that microphone condition.
- **Set of sessions:** In this case, the initial estimates of mean and variance for the current testing utterance are computed like in the previous case, but using utterances from a set of sessions of the Speecon database instead of utterances from the same session.

These two ways of doing initial estimation of mean & variance were first tested without the recursive updating (the estimates were kept constant along the given utterance). Moreover, experiments with an approximation of the second approach consisting in using the previous testing utterance to compute the estimates, were also carried out. In this case, although less data is used to do the estimation, notice that it has the advantage that it can be used in situations where the testing conditions are slowly changing.

The results are given in Table 5 and they show the performance of CT and 2-3m microphones only (similar conclusions can be obtained from the other mikes). We can see that even without using updating, the all MVNs using these estimates show better performance for the 2-3m microphone than the original recursive MVN with 0.25s look-ahead (48.77%, see Table 3). In these tests, the gap between the performance of the off-line UTT-MVN and an online MVN was reduced significantly (each of the techniques in Table 5 can be implemented without any delay) on the cost of the assumptions related to each initialization.

Mike	Set of Sessions	Current Session	
		All Utt.	Previous Utt.
CT	97.87	98.20	98.06
2-3 meters	66.37	74.02	71.23

Table 5 Non-updated MVN for initial mean & variance estimates computed from past data.

The best performance was achieved using the current session initial mean & variance estimate. In this case, a relative decrease of only 2.5% in word accuracy compared to the off-line UTT-MVN was obtained. Estimates from the set of sessions resulted to be too general obtaining the worst performance of the three approaches.

When the estimates computed from only the previous utterance were used, a relative decrease of 6% was obtained. Slight condition changes between consecutive utterances and the different quantity of data used to compute the estimates may be the reasons why estimates from the previous utterance cannot achieve the results of the current session based estimates.

As expected, when adding the recursive updating to the previous experiment, the recognition performance improves further for the all three kinds of initializations. Figure 1 shows results for several look-ahead values with the 2-3m mike. The largest improvement was obtained for the set of sessions initialization, from 66.37% to 71.85% for the 0.25s look-ahead, which shows the positive effect of the recursive updating when the initial mean & variance estimates are too general. The best absolute result for the 0.25s look-ahead, 74.73%, was obtained for the current session initialization. Figure 1 also shows the results of the recursive MVN that uses the whole current utterance to compute the initial mean & variance estimates (results obtained from Table 4). Results of this off-line technique slightly outperform the approaches based on previously recorded data, which shows the good performance of these approaches.

5. INITIAL MEAN & VARIANCE ESTIMATED FROM THE CURRENT UTTERANCE ONLY

In Section 4, where a-priori information was used for mean & variance initialization, we observed a significant improvement over the results reported in Section 3.2.1, where only the information from the current utterance was used to initialize mean & variance (assuming a maximum algorithmical delay of 0.25s). We wonder whether there is a way to improve the performance from Section 3.2.1 and still use the information from the current

utterance only. This would avoid the need of any of the assumptions mentioned in Section 4.

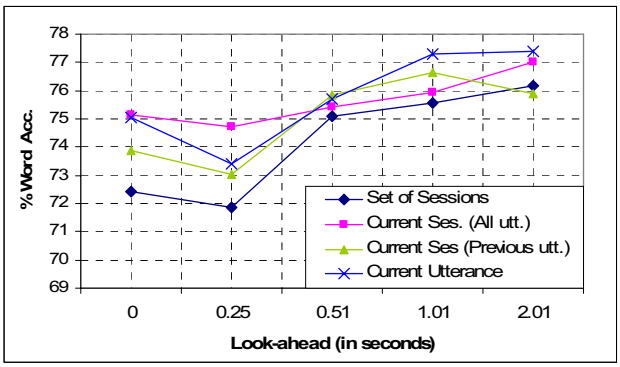

Figure 1 Word accuracies of recursive MVN comparing various initial estimates computed from past data with estimates from the whole current utterance. Scores from the test with the 2-3m mike are presented.

The results from Section 3.2.1 show that including the speech portions of signal into the initial estimation of mean & variance provides a significant improvement in the noisiest condition (the jump in the performance observed for the 2-3m microphone when using the look-ahead parameter larger than 1.25s). In addition, a mixture of speech and non-speech frames is used to compute the initial mean & variance estimates in the highly robust UTT-MVN. These observations suggest that a mixture of speech and non-speech frames should be employed for the mean & variance initialization. In order not to introduce a large delay, in the following experiments we use for the initialization the feature frames surrounding the first detected non-speech-to-speech transition. The non-speech frames until this transition are usually dropped in current ASR systems, so the only introduced delay is due to the window spanning the sequence of speech and non-speech frames used for initialization.

In our experiments, the centre of the initialization window coincides with the first detected non-speech-to-speech transition and thus, the corresponding algorithmical delay is half of the window length. Table 6 shows the results using this approach. Generally, better noisy speech performance is observed. For the max tolerable delay of 0.25ms and the 2-3m case, the performance improves from 48.77% to 64.02% when compared with Table 3.

However, in average the performance of the close-talking microphone decreased. Analysis of the results revealed the number of insertions at the first non-speech portion of each CT utterance increased as a consequence of the proposed mean & variance initialization. As a solution, frame dropping was applied and the number of insertion was effectively reduced.

To detect the first non-speech-to-speech transition, a speech activity detection (SAD) system is needed. In Table 6 we use Viterbi alignment for that purpose. We tested also a real SAD system from [12]. The good performance was preserved, which indicates a robustness of this initialization approach to the SAD decision.

	Window Length / Delay (in seconds)			
Microphone	2.0/1.0	1.0/0.5	0.5/0.25	0.2/0.1
CT	97.61	97.75	98.27	97.94
Lavalier	96.68	96.49	96.53	96.07
1 meter	95.26	94.98	94.74	93.09
2-3 meters	73.06	70.60	64.02	60.36

Table 6 Recursive MVN with initial mean & variance estimates from the first non-speech-to-speech transition

At last, when comparing Table 6 with the results in Section 4 we see it is preferable to use a-priori information to using the current utterance data only for calculating the initial mean & variance estimates (assuming max algorithmical delay of 0.25ms).

6. CONCLUSIONS

We tested several ways of implementing the mean and variance normalization technique (MVN) in the recognition of real-world noisy speech. Very good performance was obtained by the offline approach where the mean and the variance are assumed constant and are calculated previously from the whole current utterance (word accuracy of 76.02%). When however a low algorithmical delay is desirable, the mean and the variance have to be estimated/updated from segments shorter than the utterance length. In that case, a recursive MVN performs better than a segment-based MVN (we assumed 250ms as a tolerable algorithmical delay). We observed the initial estimate of mean and variance in the recursive MVN is very important; good results were obtained when using past information for this initialization (74.02%). If the past information is not available, better performance is achieved when using segments from the first non-speech-to-speech transition (64.02%) than using segments from the beginning of the utterance (48.77%). Also, we observed that the usefulness of mean and variance updating increases when the initial estimates are not representative enough for a given utterance.

7. REFERENCES

[1] S. Molau, M. Pitz, H. Ney, "Histogram Based Normalization in the Acoustic Feature Space," Proc. ASRU, 2001.

[2] O. Viikki,K. Laurila, "Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition," *Speech Communication*, vol. 25, pp. 133-147, 1998.

[3] C. -P. Chen, K. Filaliy, J. A. Bilmes, "Frontend Post-Processing and Backend Model Enhancement on the Aurora 2.0/3.0 Databases," Proc. ICSLP, 2002.

[4] O. Viikki, D. Bye, K. Laurila, "A Recursive Feature Vector Normalization Approach for Robust Speech Recognition in Noise," Proc. ICASSP, 1998.

[5] R. Haeb-Umbach, et al., "Acoustic Modeling in the Philips Hub-4 Continuous-Speech Recognition System," Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[6] C. Benitez, et al., "Robust ASR Front-End Using Spectral-Based and Discriminant Features: Experiments on the Aurora Tasks," Proc. Eurospeech, 2001.

[7] D. J. Iskra, et al., "SPEECON - Speech Databases for Consumer Devices: Database Specification and Validation," Proc. LREC, 2002.

[8] P. Pujol, et al., "Speech Recognition Experiments Speech Recognition Experiments with the SPEECON Database Using Several Robust Front-Ends," Proc. ICSLP, 2004.

[9] A. Bonafonte, et al., "RAMSES: el Sistema de Reconocimiento del Habla Continua y Gran Vocabulario Desarrollado por la UPC," Proc. VIII Jornadas de Telecom I+D, 1998.

[10] C. Nadeu, J. Hernando, M. Gorricho, "On the Decorrelation of Filter-Bank Energies in Speech Recognition," Proc. Eurospeech, 1995.

[11] F. de Wet, et al., "Additive Background Noise as a Source of non-Linear Mismatch in the Cepstral and Log-Energy Domain," *Computer Speech and Language*, vol. 19, pp. 31-54, 2005.

[12] J. Padrell, D. Macho, C. Nadeu, "Robust speech activity detection using LDA applied to FF parameters," Proc. ICASSP, 2005.