Feature versus Model Based Noise Robustness

Kris Demuynck, Xueru Zhang, Dirk Van Compernolle, Hugo Van hamme*

Dept. ESAT, Katholieke Universiteit Leuven, Leuven, Belgium

{Kris.Demuynck,Xueru.Zhang,Dirk.VanCompernolle,Hugo.Vanhamme}@esat.kuleuven.be

Abstract

Over the years, the focus in noise robust speech recognition has shifted from noise robust features to model based techniques such as parallel model combination and uncertainty decoding. In this paper, we contrast prime examples of both approaches in the context of large vocabulary recognition systems such as used for automatic audio indexing and transcription. We look at the approximations the techniques require to keep the computational load reasonable, the resulting computational cost, and the accuracy measured on the Aurora4 benchmark. The results show that a well designed feature based scheme is capable of providing recognition accuracies at least as good as the model based approaches at a substantially lower computational cost.

Index Terms: noise robustness, noise robust features, missing data theory, parallel model combination, uncertainty decoding, large vocabulary speech recognition

1. Introduction

The focus in noise robust speech recognition has gradually shifted from research on noise robust features to model based techniques such as parallel model combination [1] or uncertainty propagation by means of missing data theory [2] or uncertainty decoding [3]. Though model based approaches are based on sound principles, many approximations need to be made to keep the computational load reasonable. Even so, the resulting computational overhead is still high. Feature based approaches on the other hand exhibit low computational overhead and impose, besides the standard HMM assumptions, no additional approximations in the acoustic modelling stage. Hence, feature based systems can take into account the inter-frame correlations present in both the speech and the noise, whereas model based systems must make strong indecency assumptions between frames, states, and even Gaussian components to make their evaluation computationally tractable. The main disadvantage of feature based schemes is that the uncertainty on the resulting feature values is not taken into account during decoding, hence the importance of minimising the uncertainty.

In this work, we contrast prime examples of both feature and model based techniques in the context of large vocabulary recognition systems such as used for automatic audio indexing and transcription. The remainder of the paper is structured as follows. First, we present the reference framework under which all methods will be investigated and investigate the behaviour of some standard signal normalisation techniques. In the next two sections, we revisit model based techniques and feature based schemes. For the feature based schemes, we present noise normalisation, a preprocessing scheme that reduces the uncertainty on the resulting features. In the final section, we discuss the experimental results and draw some conclusions.

*This author is also affiliated with IBBT (Interdisciplinary Institute for Broadband Technology). This research is financed by the MIDAS project of the Nederlandse Taalunie under the STEVIN programme and the IM-PACT project BATS by IBBT.

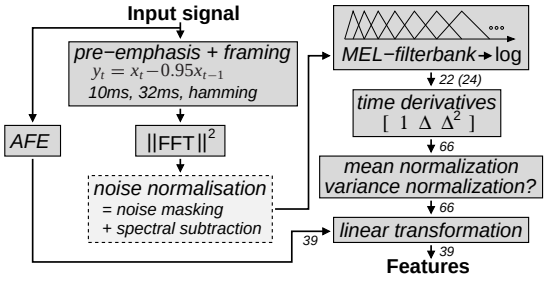


Figure 1: Preprocessing used for the baseline systems.

2. Framework & baseline systems

In this section, we present the framework under which all methods will be evaluated and investigate the behaviour of some standard signal normalisation techniques in function of the train data (clean versus multi-noise). The accuracy of the systems is measured on the Aurora4 benchmark task using the SPRAAK recognition toolkit [4]. All systems used in this paper are build around the reference preprocessing scheme shown in figure 1.

A first set of acoustic models was build on top of the ETSI advanced front-end (AFE) [5]. The AFE models act as reference for the other noise robust systems we want to investigate. The front-ends of the other systems all calculate 24 (of which 22 are used) Mel-scaled filterbank outputs (log energies) with their first and second order time derivatives. The first and last Mel-filterbank outputs are discarded since these are easily affected by external factors (microphone characteristic, aliasing filter). The resulting 66 dimensional vector is condensed to 39 dimensions using a linear transformation. The transformation consists of two parts. First, mutual information based discriminant analysis (MIDA) [6] finds the 39 dimensional sub-space which shows minimal information loss given the original 66 dimensional class distributions. The classes considered in this work are the 129 context-independent phone states. The second part decorrelates the features [6] so that the mixtures of diagonal covariance Gaussians used in the subsequent HMM deviate as little as possible from an identically configured HMM that uses full covariance Gaussians. This decorrelation step is also applied to the 39 AFE features. The automatic data-driven feature optimisation results in a good selection of features and a good conditioning of these features given the subsequent modelling. However, optimising on clean speech data gives no guarantees on the quality of the features when handling noisy speech. We therefore also trained systems using MFCC features.

Mean normalisation subtracts the long-term average from the cepstral coefficients or, equivalently, the log Mel-filterbank

id	description	clean speech training set01	01-07	08-14	multi-noise training set01	01-07	08-14	legend	
B1	MFCC	6.63	29.78	48.06	7.94	16.45	34.44	MFCC	Mel-frequency cepstral coefficients
B2	MIDA	5.77	26.91	45.35	7.21	14.72	32.16	MIDA	MIDA+decorr. on Mel-filterbank outputs
B3	MIDA-vad	5.77	27.75	45.97	7.23	14.70	32.05	AFE	ESTI advance front-end + decorr.
B4	MIDA+vn	5.83	47.61	68.55	7.81	14.21	36.87	vn	variance normalisation
B5	AFE	5.88	21.15	38.88	6.89	14.51	32.13	-vad	no leading/trailing silence removal

Table 1: Word error rates (% ins+sub+del) on the Aurora4 clean speech testset (01) and average WERs over testsets 01-07 (various noise conditions) and testset 08-14 (various noise and channel conditions) for different baseline configurations.

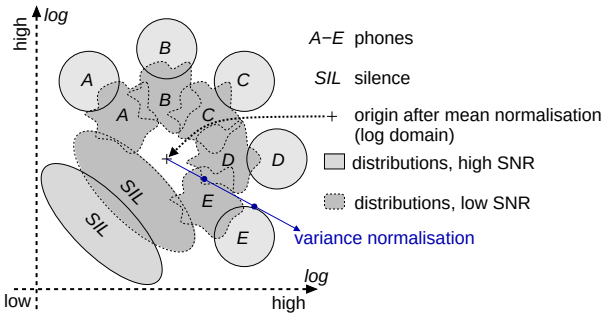


Figure 2: Conceptual model illustrating the effect of recognising noisy speech with acoustic models trained on clean speech.

outputs[1]. The presence of noise also reduces the dynamic range of the static and dynamic features. Variance normalisation has been proposed as an easy way to compensate for that.

Next to mean and variance normalisation, we also investigated the effect of adding a voice activity detector (VAD) to remove leading and trailing silence from the test sentences. This should prove helpful for systems trained on clean speech since for these systems the mismatch of the silence model under noisy testing conditions is considerable.

Table 1 lists the word error rates (WER) of the baseline systems. Comparing B1 (MFCC's) with B2 shows that automatic feature optimisation gives a 12% gain in accuracy on testdata with matching acoustic conditions (set 01 and 01-07 for clean speech and multi-noise training respectively) and even provides gains on the testsets with non-matching conditions. Variance normalisation (B4) showed to be counterproductive, especially for the clean speech systems. Begin/end-point detection (B2, B3) helps when the silence model is trained on clean speech only. The silence model in the multi-noise systems matches the testdata well enough and hence begin/end-point is not needed.

Assuming that phones are characterised by their distribution of high and low energy values (e.g. the energy dips between the formants) in the frequency domain, and assuming that the high values are unlikely to be affected by noise while the low values are easily masked by noise, the processes depicted in figure 2 are able to explain the experimental results. When using mean normalisation in the presence of noise, the origin (mean value) will be approximately halfway between the high speech values and the low noise values. This leads to a shrinkage of the acoustic space. As a result, one relies mainly on the tails of the emission probability density functions (pdf's) in the clean speech acoustic model to explain the acoustics. Assuming that the tails are less discriminative but not systematically biased, the net effect is that the relative importance of the acoustic model w.r.t. the language model will lower and some additional confusion will occur, but overall the degradation will be graceful. Variance normalisation maps the noisy values from the center back to the outer circle, and hence produces features that return decisive acoustic scores just as the clean speech features. However, since the position after variance normalisation may be incorrect due to the presence of noise, hard recognition errors can and will occur in low SNR situations.

Another interesting result was obtained when training with the union of the clean and multi-noise data. This system showed a high accuracy on both the clean speech and the noisy testsets (5.94%, 14.43% and 31.89%), showing that one can train multiple conditions into a single acoustic model. This is also in line with the conceptual model: both a clean and noisy speech model can co-exist with little overlap.

3. Model based noise robustness

In this section, we look at techniques that either modify the acoustic model or alter the way the acoustic model is evaluated.

Parallel model combination (PMC) [1] adjusts the clean speech acoustic models so that they reflect noisy speech with a noise distribution similar to the one measured on the testdata. Although the method is conceptually sound, concrete implementations require several approximations and assumptions to keep the computational overhead manageable: (1) the noise is assumed to be stationary, (2) the clean and noisy speech distributions are both modelled with mixtures of diagonal covariance Gaussians (in reality speech is reasonably well decorrelated in the cepstral domain while stationary noise is decorrelated in the spectral domain), and (3) the non-linear nature of the feature extraction is approximated with vector Taylor series (VTS).

Missing data (MD) [2] and uncertainty decoding (UD) [3, 7] work the other way around: they start from the observed noisy features and try to map them back to clean speech features. Since the presence of noise causes information loss, this reverse mapping is not deterministic but probabilistic in nature, i.e. a noisy feature vector is mapped back to a clean speech distribution. The clean speech system is then evaluated with a distribution instead of a single observation vector. This is expressed in equation 1, with $\bar{w}$ the word sequence, $\bar{y}$ the noisy speech vector, and $\bar{x}$ a possible clean speech vector.

$$P(\bar{w}|\bar{y}) = \int_{\bar{x}} p(\bar{w}|\bar{x})p(\bar{x}|\bar{y})d\bar{x} \quad (1)$$

$$\approx P(\bar{w}) \left(\prod_t \int_x \frac{p(x|s_t)}{p(x)} f(x|y_t)dx \right) \quad (2)$$

$$\sim \begin{cases} P(\bar{w}) \left(\prod_t \int_x p(x|s_t)f(x|y_t)dx \right) \\ P(\bar{w}) \left(\prod_t \max_x p(x|s_t)f(x|y_t) \right) \end{cases} \quad (3)$$

Applying HMM based Viterbi decoding, represented by the optimal state sequence s_t, leads to eqn 2. Given that the integral of

[1] Mean normalisation is a linear operation, and hence can be readily moved before other linear transformations.

the fraction cannot be expressed in a closed form, one typically relies on one of the two approximations given in eqn 3.

Both UD and MD also assume that the effect of noise (N) on the MEL-filterbank energy outputs is purely additive in nature: $Y \approx X + N$, with Y and X the noisy and clean speech filterbank outputs respectively. MD approximates the subsequent log-operator with a maximum operator ($\log(Y) \approx \max(X, N)$) whereas UD assumes normal distributions and uses VTS to propagate the uncertainty. In order to be computationally tractable, most MD and UD systems treat each Gaussian in the emission pdf's independently (yet another approximation). MD can be evaluated efficiently in the spectral domain, but this implies a significant drop in accuracy. MD in the cepstral domain is computationally expensive. UD can be sped up by clustering Gaussians so that the VTS approximation only needs to be calculated for the cluster centra [7]. Since UD and MD alter the acoustic model evaluation, techniques to speed up the Gaussian evaluation [6] (gain of a factor 20) are no longer applicable, bringing about a significant computational cost even when making further approximations such as using only a single cluster center in UD [7].

We evaluated two model based techniques. Both setups use acoustic models that are similar in size and configuration to the baseline systems. The multi-candidate model based feature enhancement (MC-MBFE) system [8] uses PMC on an ergodic speech model to estimate a clean speech distribution and propagates the uncertainty by means of multiple candidate clean speech feature vectors generated by a minimum squared error estimator. By using multiple candidates (10 in this work) instead of propagating the uncertainty in the form of a covariance matrix, the back-end decoder can still speed up the Gaussian evaluation. Consequently, MC-MBFE slows down the acoustic model evaluation with a factor 10. The back-end preprocessing is identical to baseline system B3. The second system (MD-prospect) performs MD decoding in the prospect domain [9] Working in the prospect domain lowers the computational cost for MD decoding with 40% with no loss in accuracy compared to a cepstral representation, making B1 the most comparable baseline system. Since MD-prospect modifies the acoustic model evaluation, the fast Gaussian evaluation can no longer be used, resulting in an acoustic model evaluation that is at least 20 times slower than baseline system B3.

4. Noise robust features

A common noise robust feature extraction technique is to estimate the underlying clean speech feature vectors given the noisy observation. Since robust feature extraction forgoes the uncertainty propagation done by UD and MD, one should minimise the uncertainty on the obtained estimates. AFE uses Wiener filtering to that end [5].

Instead of estimating clean speech features, one could also aim for some target signal-to-noise ratio (SNR), i.e. normalise the noise [10]. The uncertainty on the normalised features depends on how many values had to be inferred since they were obscured by noise and the possible range of the inferred values. Since adding noise ($\text{SNR}_{\text{input}} > \text{SNR}_{\text{target}}$) does not cause uncertainty on the feature values and since the range of the inferred values when decreasing the noise level is bounded by the target SNR, the resulting uncertainty will be lower than when aiming for clean speech vectors. The downside of noise normalisation is that for clean speech some information will be lost due to the injection of extra noise in order to reach the target SNR.

Feature based approaches have two main advantages over model based methods: their computational overhead is low and

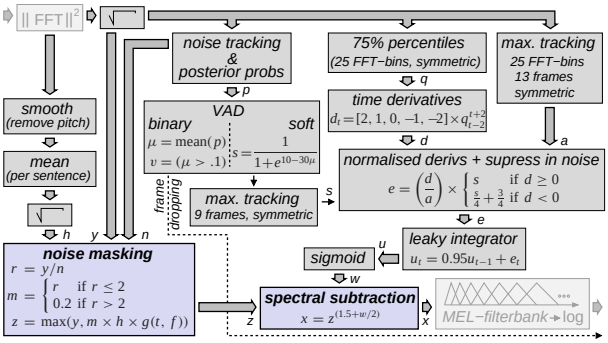


Figure 3: A noise normalisation preprocessing scheme combining noise masking with spectral subtraction.

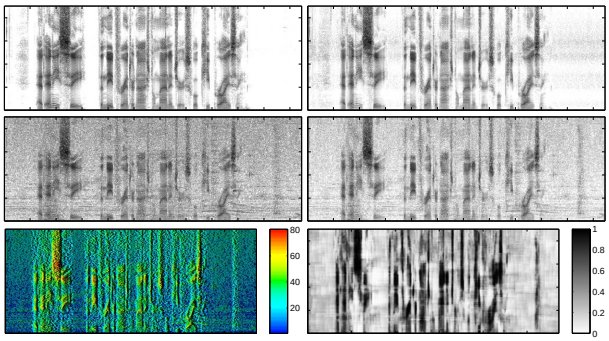


Figure 4: Top: spectrogram (0-7kHz) of a clean speech signal before (left) and after noise normalisation (right). Middle: idem for a noisy speech signal. Bottom left: spectrogram enriched with normalised time derivatives, hue and brightness correspond to energy (in dB) and time derivatives respectively. Bottom right: the output of the sigmoid function w in figure 3.

they require, besides the standard HMM assumptions, no additional approximations in the acoustic modelling stage. Hence, feature based systems can take into account the inter-frame correlations present in both the speech and the noise, whereas model based systems must make strong indecency assumptions between frames, states, and even Gaussian components to make the evaluation computationally tractable. The main disadvantage is that the uncertainty is not taken into account, hence the importance of minimising the uncertainty.

Figure 3 depicts the noise normalisation algorithm used to fill in the corresponding block in figure 1. Note that most of the processing blocks act on FFT amplitude coefficients. In essence, noise normalisation consist of two main operations: noise masking and spectral subtraction.

The noise signal m used for noise masking (increasing the noise level) is derived from the noise already present in the input signal (even for clean speech data) by dividing the immediate signal in a FFT frequency bin by the estimated noise level as returned by a noise tracker. Given that the noise tracker is only used for generating a noise signal and for VAD, a simple system with low overhead, e.g. based on minimal statistics [11], can be used. If the noise signal m were instead generated from a white noise time signal, the noise masking operation should also consider the phase of the two signals (noise and input), and should adjust the noise gain based on the level of noise already present in the input signal. By deriving the noise signal from the input signal, the noise masking can be performed with a simple $\max()$ operation. In case the input y is thought to be speech

($r > 2$), the noise injection is suppressed ($m = 0.2$).

The noise masking level is set relative to the speech level h (averaged energy per frequency bin, the energy is first smoothed in order to remove the pitch). The scale factor $g(t, f)$ is set to a constant value during testing and, except for one system (varSNR), was set to the same constant value for training.

The spectral subtraction (noise suppression) is done by applying a gain function on the log-energy values, or equivalently (see figure) by using a power less than 2 when converting the amplitude values z to energy values x. Since speech/noise classification (w) is never completely reliable, the maximum attenuation is limited to 75% of the original log energy.

The detection of speech and noise components (top right part of the figure) is based on the common on- and offset time principle used in computational auditory scene analysis [12]. Normalised time derivatives serve as simple on- and offset detectors. Operating on amplitude values instead of log-energies, renders the derivative less sensitive to noise. Extending the window for calculating the normalisation factor (inverse of the local maximum) a little bit beyond the frame range used for calculating the derivatives (13 versus 5 frames) provides additional noise suppression near speech regions. Figure 4 (bottom left) shows that adding time derivatives to a spectrogram lets the speech stand out from the noise more clearly. The "75% percentiles" block suppresses isolated on- and offsets, or otherwise said, emphasises on- and offsets that are common in the window of 25 FFT bins. By using the 75% percentile instead of the median, there is also an automatic focus on the more energetic values which are more likely to correspond to speech. The "noise suppression" component modifies the derivatives in order to bias the output of the "leaky integrator" to negative values for frames marked as silence by the VAD. The "leaky integrator" and "sigmoid" blocks convert the raw on- and offset values e into values w which serve as an estimate of the probability that an FFT bin contains speech. The time constant for the leaky integrator was chosen based on the typical rate of change (modulation frequency) observed in the spectral lines of a speech spectrogram. The other parameters were chosen based on visual inspection of the speech/noise classification values w (cf. bottom right of figure 4). Sensitivity analysis of the parameters showed that non of them are critical.

Figure 4 shows the effect of the noise normalisation on a clean and noisy speech signal. Given that the maximum noise attenuation (spectral subtraction) is limited, residual noise is still present for low SNR input signals. However, the mismatch between clean and noisy data is clearly reduced, which according to the conceptual model (figure 2) means that one will rely less on the tails of the emission pdf's.

Since input signals with a low SNR cannot be completely normalised, one could opt to train the acoustic model with variable target SNR levels (varSNR) as to make the system robust w.r.t. that variability. We trained one such system, choosing different mean values (from 0 to 0.2) for the noise gain function $g(t, f)$ for each sentence in the train data. The gain function $g(t, f)$ was furthermore set to vary slowly and randomly from the mean value both in function of t and f, trying to mimic the behaviour of the residual noise more closely.

5. Results & conclusions

Table 2 lists the results of the model and feature based schemes. Both model based systems incorporate techniques to compensate for channel mismatches [8, 9], explaining their excellent results on set 08-14. Similar techniques could be added to the feature based schemes to improve channel robustness.

system	trainset	01	01-07	08-14
MC-MBFE	clean	4.99	16.46	31.80
MD-prospect	clean	6.91	19.57	32.55
Nmask($g = 0.10$)	clean	7.08	18.61	37.38
Nmask($g = 0.14$)	clean	8.24	17.31	36.69
Nnorm($g = 0.10$)	clean	6.67	15.81	34.68
Nnorm(varSNR)	clean	6.22	14.09	32.55

Table 2: Word error rates on Aurora4 for two model based techniques (top) and a few feature based schemes (bottom).

The results show that feature based schemes are capable of providing recognition accuracies at least as good as model based approaches at substantially lower computational costs. The improvement noise masking (no spectral subtraction, denoted as Nmask) and noise normalisation (Nnorm) brings over AFE (table 1) illustrates the importance of minimising the uncertainty on the features returned by the preprocessing. The main disadvantage of noise masking, a high WER on clean speech data, can be largely mitigated by doing noise normalisation (noise masking + spectral subtraction) and by training with variable target SNR levels.

When comparing to the baseline systems, we see that the best feature based system, noise normalisation with variable target SNR levels, even rivals systems trained on (quasi) matched conditions (right-hand side of table 1).

6. References

[1] M. Gales, "Model-based techniques for noise robust speech recognition," Ph.D. dissertation, University of Cambridge, UK, Sept. 1995.

[2] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Comm.*, vol. 34, no. 3, pp. 267–285, June 2001.

[3] J. Droppo, A. Acero, and L. Deng, "Uncertainty decoding with splice for noise robust speech recognition," in *Proc. ICASSP*, Orlando, Florida, May 2002, pp. 57–60.

[4] K. Demuynck, J. Roelens, D. Van Compernolle, and P. Wambacq, "SPRAAK: An open source speech recognition and automatic annotation kit," in *Proc. ICSLP*, Brisbane, Australia, 2008, p. 495.

[5] ETSI standard doc., "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," ETSI, Tech. Rep. ES 202 050 v1.1.3, 2003.

[6] K. Demuynck, "Extracting, modelling and combining information in speech recognition," Ph.D. dissertation, K.U.Leuven, Feb. 2001.

[7] H. Liao and M. Gales, "Issues with uncertainty decoding for noise robust speech recognition," *Speech Comm.*, vol. 50, no. 4, pp. 265–277, Apr. 2008.

[8] V. Stouten, H. Van hamme, and P. Wambacq, "Model-based feature enhancement with uncertainty decoding for noise robust ASR," *Speech Comm.*, vol. 48, no. 11, pp. 1502–1514, Nov. 2006.

[9] M. Van Segbroeck and H. Van hamme, "Advances in missing feature techniques for robust large vocabulary continuous speech recognition," *IEEE Trans. on ASLP*, 2010, (accepted for publication).

[10] D. Van Compernolle, "Noise adaptation in a hidden Markov model speech recognition system," *Computer Speech and Language*, vol. 3, no. 2, pp. 151–168, 1989.

[11] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on SAP*, vol. 9, no. 5, pp. 504–512, July 2001.

[12] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Comp. Speech and Lang.*, vol. 8, pp. 297–336, 1994.