Application of noise robust MDT speech recognition on the SPEECON and SpeechDat-Car databases

J. F. Gemmeke[1], Y. Wang[2],M. Van Segbroeck[2], B. Cranen[1],H. Van hamme[2]

[1]Dept. of Linguistics, Radboud University, Nijmegen, The Netherlands.
[2]ESAT Department, Katholieke Universiteit Leuven, Belgium

{j.gemmeke, b.cranen}@let.ru.nl

{yujun.wang, maarten.vansegbroeck, hugo.vanhamme}@esat.kuleuven.be

Abstract

We show that the recognition accuracy of an MDT recognizer which performs well on artificially noisified data, deteriorates rapidly under realistic noisy conditions (using multiple microphone recordings from the SPEECON/SpeechDat-Car databases) and is outperformed by a commercially available recognizer which was trained using a multi-condition paradigm. Analysis of the recognition results indicates that the recording channels with the lowest SNRs where the MDT recognizer fails most, are also the channels which suffer most from room reverberation. Despite the channel compensation measures we took, it appears difficult to maintain the restorative power of MDT in such non-additive noise conditions.

Index Terms: Automatic Speech Recognition, Missing Data Techniques, Noise Robustness

1. Introduction

When confronted with background noise, the performance of an automatic speech recognition (ASR) engine typically degrades substantially. A practical approach to improve the noise robustness of a speech recognizer is multi-condition training [1]. Rather than training acoustic models on speech from a quiet environment only, the acoustic models are trained directly on noisy speech signals. The train set is carefully selected to reflect the multiple acoustic environments that are considered to be representative for the target application.

While often effective, especially for stationary background noise, recognition accuracies obtained with multi-condition training quickly deteriorate when the noisy environment deviates from the one that was used for training. Another disadvantage of multi-condition training is that the performance compared to matched clean conditions typically degrades. Many other approaches have been proposed to improve noise robustness [1], but for applications in which the background noise is highly variable and unpredictable, it is undesirable to model the statistics of the noise. In this paper we study Missing Data Techniques (MDTs) [2] which focus on speech properties.

The general idea behind MDT is that it is possible to estimate −prior to decoding− which speech features are reliable (i.e., dominated by speech energy) and which are unreliable (i.e., dominated by background noise). These reliability estimates are used to treat reliable and unreliable features differently and are referred to as a *spectrographic mask*. This mask can, for instance, be used to replace the unreliable features by clean speech estimates, after which recognition proceeds without further noise compensation. Thus, the acoustic models for MDT are trained on clean speech which greatly reduces the cost of data collection. Moreover, MDTs hold the promise of being more flexible in handling unexpected noises, provided the spectrographic mask is correctly estimated.

Most of the existing knowledge about the effectiveness of MDTs has been acquired using databases with noisy speech that has been constructed by artificially adding noise of various types and intensities to clean speech, in particular databases constructed in the AURORA project [3]. On these artificial data sets MDTs have proved to be able to effectively mitigate the impact of both stationary and non-stationary noise. However, very few reports exist (e.g. [4]) that describe the effectiveness of single-channel MDT recognition on real world recordings.

The goal of the current paper is to explore the effectiveness of an MDT recognizer on data recorded in real-world conditions. For this purpose we use a subset of the recordings from the SPEECON [5] (entertainment room, office and public environment) and SpeechDat-Car [6] databases and compare MDT recognition performance with that of the Nuance VOCON 3200 recognizer that has been trained with a multi-condition training approach.

The SPEECON and SpeechDat-Car databases contain simultaneous recordings from four microphones placed at different distances from the speaker, one of them being a close-talk microphone. As a result, each channel has been affected in its own specific way by the room acoustics and has an unknown amount of acoustic energy from the environment mixed in. In order to express recognition accuracy as a function of signal-to-noise ratio (SNR), we need an estimation of clean speech energy. To that end, we estimated the clean speech component in each channel using an acoustic echo canceler (AEC) which was fed with the close-talk channel as a reference signal. These clean speech estimates were also used to obtain an estimate of an ideal spectrographic mask which gives us an upper limit on the achievable MDT performance. Additionally, we conduct experiments with two blind mask estimation methods.

The rest of the paper is organized as follows. First in Section 2 we introduce MDT and the various mask estimation techniques. In Section 3 we describe the recognition task and the two recognition engines. In Section 4 we present our results and we discuss these in Section 5. Finally we present our conclusions and suggestions for future work in Section 6.

2. MDT ASR

2.1. Introduction

Under the assumption that the noise in noisy signals is additive and uncorrelated to the speech, the power spectrogram of noisy speech $\boldsymbol{Y}$, can be described as the sum of the individual power

spectrograms of clean speech $\boldsymbol{S}$ and noise $\boldsymbol{N}$, i.e., $\boldsymbol{Y} = \boldsymbol{S} + \boldsymbol{N}$. Elements of $\boldsymbol{Y}$ that predominantly contain speech or noise energy are distinguished by a spectrographic mask $\boldsymbol{M}$:

$$M(k,t) = \begin{cases} 1 \stackrel{def}{=} \text{reliable} & \text{if } 10\log_{10}[S(k,t)/N(k,t)] > \theta \\ 0 \stackrel{def}{=} \text{unreliable} & \text{otherwise} \end{cases} \quad (1)$$

with frequency band k ($1 \leq k \leq K$), time frame t ($1 \leq t \leq T$) and constant threshold θ (in dB).

Using log-compressed features, the reliable noisy speech features can be used directly as estimates of the clean speech features [2]. For the unreliable features we provide clean speech estimates during decoding [7]. In experiments with artificially added noise, the so-called *oracle masks* can be computed directly using Eq. 1. In realistic situations, however, the masks must be estimated [8]. In the next subsections we describe the mask estimation methods employed in this work.

2.2. VQ missing data masks

As a first approach to estimate spectrographic masks, we employ the Vector Quantization (VQ) strategy proposed in [9]. Here, the key idea is to estimate masks by making only weak assumptions about the noise, while relying on a strong model for the speech. The speech model uses a-priori knowledge about how the human voice manifests itself in terms of harmonicity, voicing, onset, etc. We seek to recover the original clean speech vector from a set of stored codewords by minimizing a cost function that is robust against additive noise corruptions. To compensate for linear channel distortions, the VQ-system self-adjusts the codebook containing the codewords to the channel during online recognition. Finally, the spectrographic VQ-based mask is estimated by thresholding the ratio of speech and noise power estimates using Eq. 1. The VQ-codebook was trained on features extracted from a subset of the training data (containing close-talk channel speech) and the number of codebook entries was limited to 1000. Recognition tests on the test set using a large interval of threshold values revealed that the threshold setting was not very sensitive. The (optimal) results presented in this work were obtained with $\theta = 9$ dB.

2.3. SVM missing data masks

A different approach to mask estimation is to use machine learning to directly classify each feature as either reliable or unreliable. Inspired by the Bayesian classification approach in [10], we created a mask estimation procedure using several independent Support Vector Machine (SVM) classifiers. We trained a separate SVM classifier for each of the $K = 22$ Mel frequency bands using LIBSVM [11]. Each classifier used the same set of single-frame-based features consisting of the 'Subband Energy to Subband Noise Floor Ratio' and 'Flatness' features derived from the noisy Mel-spectral features described in [10], as well as all features used for the VQ-based mask estimation described in the previous Section. The SVMs were trained on 2000 frames (amounting to 20 seconds of speech) randomly extracted from the Aurora-4 multi-condition training set [12]. Reliability labels used in training were obtained from the oracle mask, derived by using the (available) clean speech and noise sources in Eq. 1 with $\theta = -3$ dB (cf. [9]). We used an RBF-kernel and hyperparameters were optimized by doing 5-fold cross validation on the training set.

2.4. Semi-oracle mask

With the single microphone recordings used in this paper, the clean speech and noise components are not known exactly. As a consequence, oracle masks for obtaining an estimate of the upper bound on recognition performance with MDT cannot be computed. In order to approximate the oracle mask, we use an acoustic echo canceler (AEC) to predict the clean speech component in the far-talk channels from the close-talk channel.

Put in the terminology of literature on AEC: the close-talk channel serves as the far-end signal, while the noisy speech in the far-talk channel takes the role of the near-end signal. The AEC then estimates an adaptive Finite Impulse Response (FIR) filter which approximates the near-end signal by filtering the far-end signal. The FIR filter models the acoustic path from the close-talk microphone to the far-talk microphone. Hence, filtering the close-talk signal projects the high-SNR signal to the far-talk microphone and serves as an estimate of the clean speech component in the far-talk channel. Subtracting it from the real far-talk signal yields an estimate of the noise picked up by the far-talk microphone. Thresholding the ratio of speech and noise power estimates using Eq. 1 we obtain the semi-oracle mask.

Because the noise in the far-talk channel is permanently present, the AEC works in double-talk conditions. This leads to biased estimates, poor convergence or even divergence of the FIR filter coefficients. To mitigate the bias caused by the double-talk condition, we selected the PEM-AFROW algorithm [13], which estimates pre-whitening filters along with the FIR filter (with an impulse response length of 20ms).

Recognition tests using a large interval of threshold values yielded optimal results with $\theta = 6$ dB for the car and public hall environments and $\theta = -1$ dB for the entertainment room and office environments.

3. Experimental Setup

3.1. Speech material

For our recognition experiments, we used a subset of the isolated word data in the Flemish part of the SPEECON [5] and the SpeechDat-Car [6] databases. We selected speech recorded in four environments (office, public hall, entertainment room and car) using four microphones with different directional characteristics which were placed at various distances (up to 1.5 m) from the speaker. Channel #1 always pertains to a close-talk microphone.

The complete test set contains 157k words. In order to reduce the computation time needed for recognition, we constructed a subset containing a balanced mixture of SNR conditions: First, we estimated the clean speech and noise signals using the AEC as described in Section 2.4. Next, after dividing the signal into silence and speech parts using an energy-based voice activity detection algorithm, the SNR per utterance was estimated, using the average frame energy of the speech-only frames of the estimated clean speech signal and of the noise-only frames of the estimated noise signal.

Using these SNR estimates we created 6 SNR subsets, each with a 5 dB binwidth, spanning a 0 dB to 30 dB range. The SNR subsets were filled by randomly selecting 700 utterances per SNR subset, ensuring a uniform word occurrence. The number of utterances per channel is not evenly distributed per SNR bin: While there are some differences per noise environment, generally speaking, the highest SNR bins mostly contain utterances from channel #1 while the lowest SNR bins mostly contain channel #4 speech.

3.2. Recognizers

In our experiments, we use the ESAT MDT recognizer and the VOCON 3200 recognizer (version 2.6) from Nuance Commu-

nications. The VOCON 3200 ASR engine is a small-footprint engine, using MFCC based features and HMM models. It uses the multi-condition training approach in tandem with speech enhancement techniques to cope with stationary or slowly varying background noise.

The details of the ESAT MDT-recognizer can be found in [7, 14]. In a nutshell, 22-channel MEL filter bank spectra and their derivatives are transformed to the PROSPECT domain [7], where clean speech estimates are obtained during decoding. The derivative features are imputed using separate masks, derived from the static mask as in [15]. Convolutional noise is compensated by updating an initial channel estimate through maximization of the log-likelihood of the best-scoring state sequence of a recognized utterance [14].

The Flemish acoustic model was trained on 60 hours of close-talk channel speech data in the SPEECON database, resulting in a triphone model containing 2534 tied states and 28917 tied mixtures.

4. Results

Figs. 1a through 1d show word error rates (WER) for the isolated word recognition task in the four different acoustic environments. In all environments, the MDT system is outperformed by VOCON3200 at SNRs below 20 dB. Above 20 dB SNR, all recognition accuracies are comparable. The one exception is the WER obtained with semi-oracle mask estimation in the car environment (Fig. 1a), which shows a higher WER (3.5% against 1.5%) in the 25-30 dB SNR range.

Except for the car environment, recognition performance with the MDT recognizer deteriorates very rapidly with decreasing SNR. Even at a moderate 15-20 dB SNR range, the WER is generally too high to be usable in a practical application. Despite the fact that the semi-oracle mask does perform better than the masks estimated solely on the basis of a single recording channel, WERs are still disappointingly high, especially at the lowest SNR ranges.

Differences between the results obtained with VQ and SVM based mask estimation methods are small in the entertainment room and public hall environments (Figs. 1b, 1d). In the car and office environments (Figs. 1a, 1c) the differences are more substantial, up to $\approx 20\%$ absolute WER in the 0-5 dB SNR range.

5. Discussion

Although our MDT recognizer has shown acceptable performance on artificially noisified data at low SNRs (e.g., [9]), it apparently fails do to so on real, noisy speech recordings. Also, the differences in performance between the multi-condition trained VOCON recognizer and our MDT recognizers increase much more rapidly with decreasing SNR than expected on the basis of previous experiences with artificially noisified data.

Comparing the recognition accuracies for the estimated VQ and SVM masks with those for the semi-oracle mask show that the recognition errors of our MDT recognizer are at least partly due to mask estimation errors. For both masks the difference in performance with the semi-oracle mask is quite substantial. Moreover, although the SVM-based mask consistently performs better than the VQ-based mask, the WER differences achievable with the two estimated masks are small.

The most striking result in Fig. 1 is the relatively high WER obtained at low SNRs with the semi-oracle mask. Although MDT is expected to perform somewhat worse than a multi-condition trained recognizer, a more thorough investigation why the difference with the VOCON results is so large at low SNRs is in order.

A first explanation is that the semi-oracle mask is derived under the assumption that the close-talk signal can be considered as 'clean' speech. However, the close-talk microphone may also pick up environmental noise or non-speech speaker sounds. Clearly, our semi-oracle mask is not the ideal mask (as opposed to the oracle mask in Aurora experiments). If the close-talk signal also contains noise, the clean speech and noise estimated from the AEC-procedure (cf. Section 2.4) may be wrong, causing errors in the semi-oracle mask.

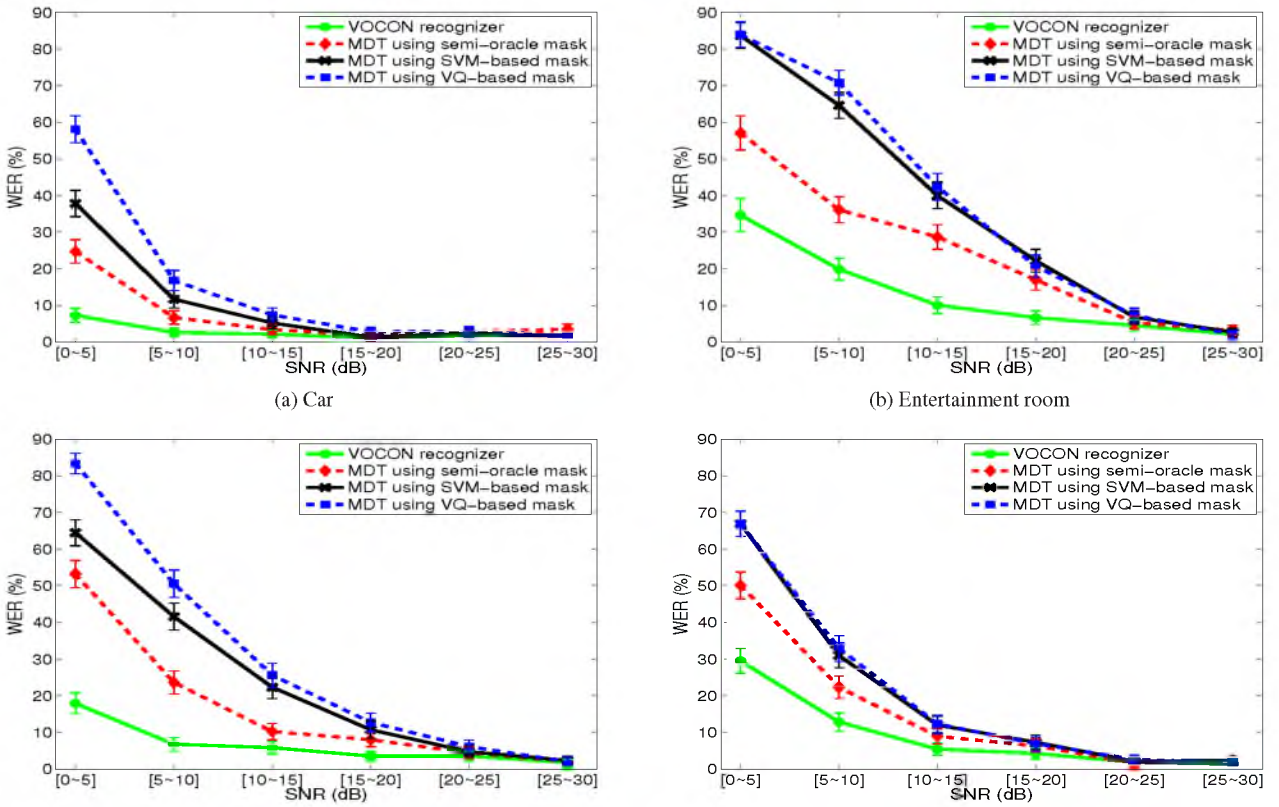
Another explanation for the higher WERs in the lower SNR-ranges relates to the fact that they originate from different recording channels than those for the higher SNR ranges. As mentioned in Section 3.1, with decreasing SNR, more and more utterances are taken from channels #3 and #4. In fact, for the lowest SNRs of the two worst-performing environments, viz. entertainment room and office, virtually all utterances stem from channel #4. Since these utterances are recorded via an omni-directional microphone, they are not only likely to contain more noise energy, giving rise to lower SNRs, but they probably also contain more reverberation effects. In case of reverberation, it is conceivable that the duration of the impulse response of the FIR filter (20 ms) in the AEC is too short.

Detailed analysis of WER as a function of both channel and SNR (not shown here) revealed that indeed, even channel #4 utterances with a high SNR(-estimate) had a high WER. The underlying reason is probably that the (estimated) clean speech in the far-talk channel does not fit the acoustic models very well which are trained on the utterances of channel #1. Regardless whether this mismatch is caused by a failing AEC or something else, the assumption that 'reliable' time-frequency cells are uncorrupted and therefore require no further processing is apparently violated. The VOCON recognizer, in contrast, employs multi-condition acoustic models which implicitly do capture all inter-channel differences.

6. Conclusions and future work

Reporting WER as a function of SNR, as is customary for experiments on artificially noisified data, we observed recognition performance to drop at a much faster rate with decreasing SNR than expected on the basis of experiments with Aurora databases. Moreover, at low SNRs, we were not able to achieve WERs with our MDT recognizer that are comparable to those of a commercial VOCON3200 recognizer which used a multi-condition training paradigm in combination with a traditional denoising front-end, not even when (semi) oracle masks were used. Further analysis showed that the SNR range categories differed not only with respect to the additive noise energy but also with respect to specific recording channel properties (like microphone type and reverberation).

The most important lesson to be learned from the current experiments is that it is not straightforward to transfer the MDT paradigm from an artificial framework like Aurora to a more real-world situation. Clearly, many of the procedures that appeared to be successful on artificially noisified databases are much less successful on single channel, far-talk signals. Using the close-talk signal as an auxiliary channel to estimate a semi-oracle mask already appeared to be a non-trivial challenge, let alone the VQ and SVM mask estimation procedures which use only single-channel data. For MDT-ASR to be effective in real world environments, not only the mask estimation procedures need improvement, but also training procedures must be developed that yield (clean speech) acoustic models which are less



(a) Car

(b) Entertainment room

(c) Office

(d) Public hall

Figure 1: WER of the isolated words recognition task recorded in a four different environments. These results were obtained with the VOCON 3200 recognizer and the MDT based recognizer using respectively a semi-oracle, SVM-based and VQ-based missing data mask estimation. The vertical bars around the data points indicate 95% confidence intervals.

sensitive to idiosyncracies of the recording channel. Therefore, our future work consists of experimenting with multi-channel acoustic models in the MDT framework. Also, we will investigate the effectiveness of measures against reverberation [16].

7. Acknowledgments

This research is financed by the MIDAS project of the Nederlandse Taalunie under the STEVIN programme. The research of Maarten Van Segbroeck was financed by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

8. References

[1] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall, 2001.

[2] B. Raj and R. Stern, "Missing-feature approaches in speech recognition," *Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, 2005.

[3] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. of ISCA ASR2000 Workshop*, 2000, pp. 181–188.

[4] U. Remes, K. J. Palomäki, and M. Kurimo, "Missing feature reconstruction and acoustic model adaptation combined for large vocabulary continuous speech recognition," in *Proc. of EUSIPCO*, 2008.

[5] D. Iskra, B. Grosskopf, K. Marasek, H. V. D. Heuvel, F. Diehl, and A. Kiessling, "Speecon - speech databases for consumer devices: Database specification and validation," in *Proc. of LREC, year = 2002, pages = 329–333.*

[6] H. V. D. Heuvel, J. Boudy, R. Comeyne, and M. N. Communications, "The speechdat-car multilingual speech databases for in-car applciations," in *Proceedings of the European Conference on Speech Communication and Technology*, 1999, pp. 2279–2282.

[7] H. Van hamme, "Prospect features and their application to missing data techniques for robust speech recognition," in *Proc. of INTERSPEECH*, 2004, pp. 101–104.

[8] C. Cerisara, S. Demange, and J.-P. Haton, "On noise masking for automatic missing data speech recognition: A survey and discussion," *Computer, Speech and Language*, vol. 21, no. 3, pp. 443–457, 2007.

[9] M. Van Segbroeck and H. Van hamme, "Vector-Quantization based mask estimation for missing data automatic speech recognition," in *Proc. ICSLP*, 2007, pp. 910–913.

[10] M. Seltzer, B. Raj, and R. Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, pp. 379–393, 2004.

[11] C. Chang and C. Lin, "Libsvm: a library for support vector machines," 2001.

[12] N. Parihar and J. Picone, "An analysis of the aurora large vocabulary evaluation," in *Proc. of Eurospeech*, 2003, pp. 337–340.

[13] T. van Waterschoot, G. Rombouts, P. Verhoeve, and M. Moonen, "Double-talk-robust prediction error identification algorithms for acoustic echo cancellation," *IEEE Transactions on Signal Processing*, vol. 55, no. 3, pp. 846–858, 2007.

[14] M. Van Segbroeck and H. Van hamme, "Handling convolutional noise in missing data automatic speech recognition," in *Proc. ICSLP*, 2006, pp. 2526–2565.

[15] H. Van hamme, "Handling time-derivative features in a missing data framework for robust automatic speech recognition," in *Proc. of ICASSP*, 2006.

[16] K. J. Palomki, G. J. Brown, and J. Barker, "Missing data speech recognition in reverberant conditions," in *in Proc. of ICASSP*, 2002, pp. 65–68.