ON THE RELATION BETWEEN STATISTICAL PROPERTIES OF SPECTROGRAPHIC MASKS AND RECOGNITION ACCURACY

J. F. Gemmeke and B. Cranen and L. ten Bosch
Dept. of Linguistics
Radboud University
P.O. Box 9103, NL-6500 HD
Nijmegen, The Netherlands
email: {J.Gemmeke, B.Cranen, L.tenBosch}@let.ru.nl

ABSTRACT
Missing Data Techniques (MDT) can significantly improve the accuracy of automatic speech recognition (ASR) for speech corrupted by background noise. The increase in recognition accuracy obtained using MDT is largely dependent on the estimation of spectrographic masks used to distinguish speech from noise. We present an analysis technique which enables us to compare two mask estimation techniques. By contrasting a sound-class independent and a sound-class dependent distance measure, we show that we can directly relate differences between masks to their difference in recognition accuracy using the sound-class dependent distance measure. Experiments on AURORA-2 using an oracle mask and an estimated mask show that modifying the estimated mask in order to reduce the statistical differences with the oracle mask leads to an increase in word recognition accuracy.

KEY WORDS
Speech Processing, Time-Frequency Signal Analysis, Robustness, Missing Data Techniques

1 Introduction

Automatic speech recognition (ASR) suffers from reduced recognition accuracies when speech is corrupted by background noise. The application of Missing Data Techniques (MDT) [1, 2, 3] can significantly improve the noise robustness, both for stationary and non-stationary noise. In MDT all time-frequency 'cells' of a spectrographic representation are labeled 'reliable' or 'unreliable'(missing). In an unreliable cell the noise energy dominates, while in a reliable cell the speech energy exceeds the noise. During recognition the unreliable parts of the acoustic vectors can be restored (feature vector imputation [4, 5]), or the decoder can be modified so that it can deal with missing data directly (marginalization [2]).

In experiments with artificially added noise, reliable/unreliable decisions can be made using knowledge about the corrupting noise and the clean speech signal; this results in so-called 'oracle' masks. In realistic situations, however, the masks must be estimated. Many different estimation techniques have been proposed, such as SNR based estimators [6], methods that focus on speech characteristics, e.g. harmonicity based SNR estimation [7], mask estimation by means of Bayesian classifiers [8, 9] and masks composed of spectro-temporal fragments [10]. We refer the reader to [11] and the references therein for a more complete overview of mask estimation techniques. In practice, estimated masks always yield lower recognition accuracies than oracle masks. This makes the oracle mask a good starting point to explore which properties make it superior.

According to [8] it is not possible to predict recognition accuracy by a direct comparison of an estimated mask and the oracle mask. However, in this paper we show that by selecting a proper distance function, we *can* relate differences between statistical properties of an estimated mask and the oracle mask to the difference in recognition accuracy. We also show that modifying the estimated mask in order to reduce the distance between the estimated mask and the oracle mask leads to an increased recognition accuracy of the estimated mask. The distance measure introduced in this paper will make future research in missing feature techniques more efficient, because it is easier to improve mask estimation techniques when it is clear *where* they are deficient. In addition, the time needed for experiments is shortened if the number of recognition runs can be reduced.

The statistical representation that we propose to characterize the differences between two masks is the proportion of frames in which each individual frequency band is considered as reliable, denoted as the *mask frame average* (***MFA***). A complete ***MFA*** is a vector with for each frequency band the probability that the frequency band is reliable.

Speech sounds are characterized by the energy distribution over frequency. Since lower energy levels are more likely to be obscured by background noise, it is reasonable to assume that individual (classes of) speech sounds have their own set of frequency bands that are crucial for identification. Therefore, we will investigate differences between masks corresponding to classes of sounds with similar spectral envelopes. Thus, ***MFA***'s are constructed by selecting mask frames that pertain to the same phone. Furthermore, we compare two approaches in defining the distance between ***MFA***'s. The first approach includes all fre-

quency elements of *MFA*'s. In the second approach we restrict the distance measure to the subset of frequency bands that are most characteristic for a specific speech sound.

In this paper we illustrate our approach by investigating the difference between an oracle mask and the harmonicity mask [7]. We are, however, confident that the results will generalize to most, if not all, other types of masks that have been proposed in previous research. We conduct our analysis and experiments using the AURORA-2 continuous digit recognition task, because this allows us to compute oracle masks.

The remainder of this paper is organized as follows. In Section 2 we describe our analysis framework. In Section 3 we present an analysis which relates the differences between *MFA*'s to differences in recognition accuracy. In Section 4 we explore the effects of decreasing the differences between *MFA*'s on the recognition accuracy. We discuss our general findings in Section 5. We summarize our findings and discuss future work in Section 6.

2 Method

2.1 Frame-based labeling

In order to be able to focus on the acoustic mismatch caused by the added noise, we use a free-phone recognition system - i.e., one without dictionary or grammar - to label individual frames. By doing so, we avoid the bias that would be introduced by the phonotactic and word sequence constraints of a regular recognizer. To provide a reference transcription at frame level we first labeled all frames in the AURORA-2 corpus by means of forced alignment of the clean speech with a canonical phone transcription. These reference labels were then compared with the labels assigned by free phone recognizers that decoded the noisy speech.

2.2 Spectrographic mask estimation

The 'oracle' mask is constructed by comparing the log energy spectra of the clean speech S and the added noise N. For the reliability of a time-frequency cell we write:

$$M(k,j) = \begin{cases} 1 \stackrel{def}{=} \text{reliable} & S(k,j) \geq (N(k,j) - \theta) \\ 0 \stackrel{def}{=} \text{unreliable} & \text{otherwise} \end{cases} \quad (1)$$

with k the frequency band, j the time-frame, and $\theta = 3$ dB a fixed mask threshold.

The harmonicity mask [7] decomposes the noisy speech signal in a harmonic and random part. It then estimates the local energy of speech and noise by thresholding the ratio between the harmonic and random part analogous to Eq.1. Parameter settings for the harmonicity mask were the same as in [7]. For both the oracle and the harmonicity mask delta and delta-delta coefficients were constructed using the procedure described in [12].

2.3 Speech recognition engine setup

We used the ESAT speech recognizer [13]. Acoustic feature vectors consisted of mel frequency power spectra (18 bands with center frequencies starting at 100 Hz, as well as first and second derivatives, i.e. 54 coefficients in total), which are then converted to 54 PROSPECT features [14]. Unreliable features are replaced by estimated values using maximum likelihood per Gaussian-based imputation [14]. We trained 195 context-dependent three-state phone models using clean speech. During decoding the context-dependent phones were mapped on 21 phone labels. To provide a reference transcription at frame level we used ESAT's Viterbi alignment package (Vitalign).

2.4 Mask Frame Averages (MFA)

Given a collection of J frames we consider the corresponding mask $M(k,j)$ with frequency band k and frame j ($1 \leq j \leq J$). For each set of J frames we calculated the mask frame average vector *MFA* with components:

$$\text{MFA}(k) = \frac{1}{J} \sum_{j=1}^{J} M(k,j) \quad (2)$$

3 Analysis

In this section we first explain the ways in which we calculate *MFA* for different subsets of the frames. Next, we introduce two distance measures for comparing *MFA* values of corresponding subsets. Results obtained with these distance measures are shown in subsection 3.4 and discussed in subsection 3.5.

All experiments were performed with test set A of the AURORA-2 continuous digit recognition task.

3.1 Frame selection for MFA construction

To construct sets from which potentially meaningful *MFA*'s can be calculated, we constructed a frame-based database using the results of two recognition tasks, one with the oracle mask and one with the harmonicity mask. Every frame in the database referred to the acoustic vector, the reference transcription phone label, the oracle mask vector, the phone label obtained using the oracle mask, the harmonicity mask vector and the phone label obtained using the harmonicity mask. This database allowed us to compare phone labels as a function of sound-class and mask type. From the frames in the database subsets were selected using the following procedure (cf. Fig. 1):

- Create subsets of frames that were labeled incorrectly (when compared to the reference transcription) using a harmonicity mask, but correctly using an oracle mask (marked *HIOC*) and frames that were labeled correctly with both mask types (called *HCOC*). Those

frames that were labeled incorrectly with both mask types (*HIOI*) were not used. We found no frames that were labeled correctly using a harmonicity mask, but incorrectly using an oracle mask (*HCOI*).

- Subdivide the *HIOC* and *HCOC* subsets on the basis of the reference transcription phone labels. This allows us to apply a sound-class (phone) dependent distance measure. We shall denote such subsubsets as $HIOC_{label}$ and $HCOC_{label}$ respectively, where 'label' stands for 'phone (class)'.

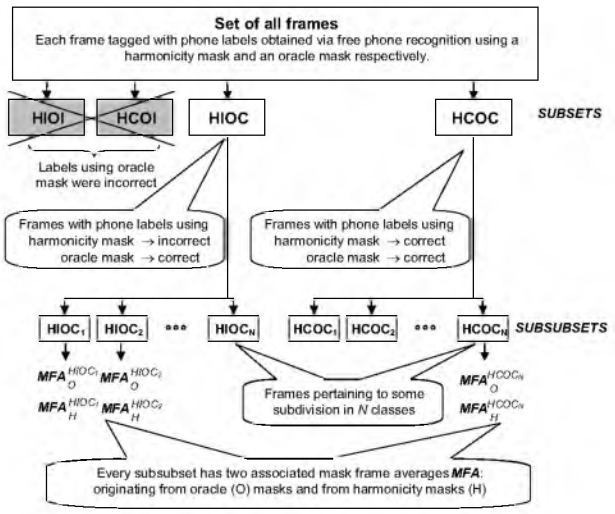

Figure 1. The hierarchy of sets, subsets and subsubsets resulting in mask frame averages.

For each of these subsubsets we calculate $\mathbf{MFA}^{S_l}$ specified by subset S and label l.

3.2 Relating MFA's differences to differences in recognition accuracy

A difference between the oracle mask and the harmonicity mask which might be related to a difference in phone labeling accuracy should become apparent by studying the difference between $\boldsymbol{MFA}_O$ and $\boldsymbol{MFA}_H$ constructed from the oracle and the harmonicity mask:

$$\mathrm{DMFA}^{S_l} = D(\mathbf{MFA}_O^{S_l}, \mathbf{MFA}_H^{S_l}) \quad (3)$$

with D some distance function. The distance functions are described in subsection 3.3.

Since in the *HCOC* subset both mask types led to correct phone labeling, while in the *HIOC* subset only the oracle mask led to correct phone labeling, a suitably chosen distance function should result in different distances between the oracle and harmonicity ***MFA***'s in the two subsets. For every label l we express this difference as:

$$\Delta\mathrm{DMFA}^l = \mathrm{DMFA}^{HIOC_l} - \mathrm{DMFA}^{HCOC_l} \quad (4)$$

Since recognition accuracy is always higher for oracle masks than for estimated masks, we expect DMFA to be smaller in the (*HCOC*) subsets than in the (*HIOC*) subsets. In other words, we aim for a distance function Eq. 3 that yields positive values for $\Delta DMFA$ for all labels l. Because every frequency band of a ***MFA*** can be considered to be drawn from a binomial distribution, we calculate confidence intervals for ***MFA***, $DMFA$ and $\Delta DMFA$.

3.3 Distance between MFA's

As a first approach for the distance function we used the L_1 distance:

$$D(\mathbf{MFA}_O, \mathbf{MFA}_H) = \sum_{\text{all } k} |\mathrm{MFA}_O(k) - \mathrm{MFA}_H(k)| \quad (5)$$

with frequency bands k. As a second approach we propose a distance measure which only considers a *selection* of frequency bands $\mathcal{K}$ which we assume to be most crucial for correct recognition of a sound class with label l:

$$D(\mathbf{MFA}_O, \mathbf{MFA}_H) = \sum_{k \in \mathcal{K}} |\mathrm{MFA}_O(k) - \mathrm{MFA}_H(k)| \quad (6)$$

with frequency bands k. $\mathcal{K}$ is a set containing the indices of K frequency bands with the highest a-priori likelihood of being reliable as determined from the oracle ***MFA***. In this paper we used $K = 5$ which was empirically found to be optimal.

3.4 Analysis results

Testset A of the AURORA-2 corpus is divided into 24 sets, i.e., four noise types and six SNR values, SNR= $20, 15, 10, 5, 0, -5$ dB. The clean speech in testset A of the AURORA-2 corpus is not used. $\Delta DMFA^l$ is calculated for every noise condition and for every phone label l. By averaging over all phones we obtained 24 $\Delta DMFA$ values, one value for each testset in testset a of the corpus. This procedure was followed for both distance measures. Note that averaging over all phones is not equivalent to calculating $\Delta DMFA$ over all phones directly since the distance measure in Eq. 6 is phone-dependent.

Example mask frame averages ***MFA*** for the phone */ah/* are shown in Fig. 2. The subsubset $HCOC_{ah}$ consisted of 9965 frames and the $HIOC_{ah}$ subsubset of 1171 frames. The difference between the ***MFA***'s of oracle and harmonicity masks for all phone labels using Eq. 5 is shown in Fig. 4 for SNR= 5 dB. Fig. 6 shows the distance when Eq. 6 is used for SNR= 5 dB. Fig. 3 shows in a histogram the average $\Delta DMFA$ for each noise and SNR subset in testset A of the corpus. The error bars in Figs. 4 and 6 denote standard deviation.

3.5 Discussion

The results show that the reliability of some frequency bands is more important for correct recognition, or in this case, correct frame labeling. This becomes already apparent when studying individual ***MFA***'s as in Fig. 2: Figs. 2(a), 2(b) and 2(c) all represent ***MFA***'s corresponding to correct phone labeling, but they differ greatly. The results in Figs. 4 and 6 also differ due to different selections of frequency bands.

The positive values in Fig. 6(c) show that $\Delta DMFA$ computed with Eq. 6 predicts the difference in labeling accuracy between oracle and harmonicity masks. This is in contrast to the $\Delta DMFA$ values observed in Fig. 4(c), created using the L_1 distance measure (Eq. 5), showing values close to zero for the majority of the phone labels. The same contrast can be found by observing that the two distance measures form two distinct classes when plotted as a histogram in Fig. 3. This shows that the differences between these distance measures are consistent for all SNR's and noise types in testset A of the Aurora-2 database. The claim that it is not possible to predict recognition accuracy by a direct comparison of an estimated mask and the oracle mask in [8] is supported when using the L_1 distance function (Eq. 5) but contradicted using the proposed sound-class dependent distance function (Eq. 6). Therefore, it seems that differences between masks *can* predict differences in accuracy, provided that the differences between the masks are expressed using subset-averaging in the form of ***MFA***'s and a sound-class dependent distance function.

A closer examination of Fig. 6(c) shows that the phone labels for which $\Delta DMFA$ fails to predict recognition accuracy are */f/*, */k/*, */s/*, */t/*, */th/*, */v/* and */z/*. The plosives */t/* and */k/* are characterized by their non-stationarity. It is unsurprising that a distance measure which only takes the reliability of static features into account is not very successful for this type of sounds. A study of the masks of delta and delta-delta coefficients might therefore yield better results for these phone labels. The phones */f/*, */s/*, */th/* and */z/* are characterized by somewhat diffuse spectral maxima and (except for */s/*) a relatively low intensity. The diffuse spectrum may affect the prediction which frequency bands are particularly important for recognition. The phone */v/* is voiced as opposed to */f/*, possibly explaining the slightly better result.

4 Biasing estimated masks towards the MFA's of the oracle mask

So far we only have shown that there is a relation between *DMFA* and recognition accuracy. However, it remains to be shown that a modified estimated mask, with reduced *DMFA*, will indeed result in higher recognition accuracy.

4.1 Mask manipulation

An approach to reduce *DMFA* is by modifying the harmonicity mask values in such a way that it's ***MFA***'s more closely resemble oracle ***MFA***'s. We do this by randomly switching selected time-frequency cells from reliable to unreliable and vice versa. We bias the probability of these random changes depending on the distance between the harmonicity and oracle ***MFA***'s. Using the procedure described below we obtain a harmonicity mask that is statistically more similar to the oracle mask:

- For every frame in the harmonicity mask, select the oracle and harmonicity ***MFA*** depending on the sound class of that frame.
- For every frequency band k of a set of $\mathcal{K}$ frequency bands, consider the difference:

  $$T_k = \text{MFA}_O(k) - \text{MFA}_H(k), k \in \mathcal{K}$$

  It follows that $T_k \in [-1, 1]$.
- For every frequency band k in $\mathcal{K}$, use a random number $R \in [0, 1]$ as the probability that the mask value of this time-frequency cell changes using the difference T_k as a threshold:

  $T_k > 0$: If $R < |T_k|$ 'unreliable' changes to 'reliable'
  $T_k < 0$: If $R < |T_k|$ 'reliable' changes 'unreliable'

In analogy with Section 3, we explore two selection criteria for the frequency bands k:

- $\mathcal{K}$ containing all frequency bands in the time-frequency representation.
- $\mathcal{K}$ containing the indices of K frequency bands with the highest a-priori likelihood of being reliable as described in subsection 3.2. As before, we used $K = 5$.

Using this procedure we bias the ***MFA***'s of the harmonicity mask toward the ***MFA***'s of the oracle mask since a larger distance between the two ***MFA***'s increases the probability that the mask values changes. This approach reduces *DMFA* without taking individual time-frequency reliability scores into account thus avoiding the trivial solution of an exact matching oracle mask.

We performed phone and word recognition using the newly constructed masks. Phone recognition was done with the same triphone models as used for the free-phone decoder. Word recognition was performed using an addition phone bi-gram "language model".

4.2 Experimental results

Fig. 5 shows recognition accuracies with the post-processed harmonicity mask obtained with the same free phone recognizer used in experiment 1.

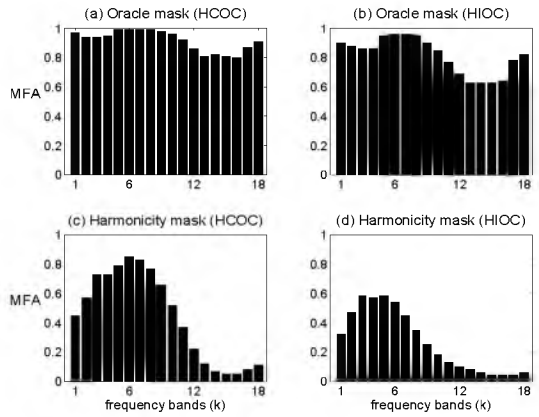

Figure 2. Example. The Mask Frame Average (***MFA***) of frames in the *HCOC* subset (a) and (c), and the *HIOC* subset (b) and (d). The sound class used to select the subsubset was the phone label */ah/*. The ***MFA*** shows for every frequency band the fraction of frames that were considered reliable

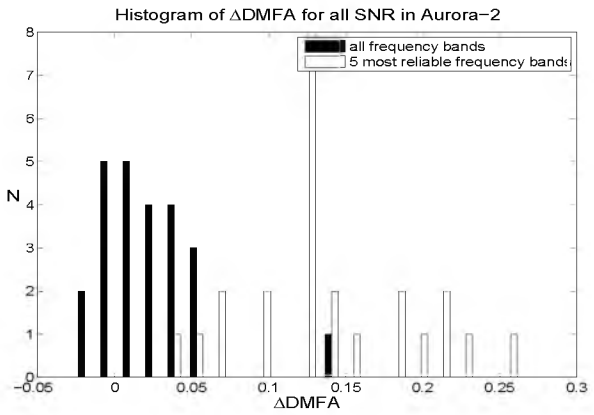

Figure 3. A histogram with phone averaged $\Delta DMFA$ values obtained by analyzing testset **A** of the AURORA-2 corpus. The corpus was divided in 4 noise types for every SNR. SNR values were $20, 15, 10, 5, 0, -5$ dB. This resulted in 24 analysis tasks. The y-axis shows the number N of analysis tasks resulting in a certain $\Delta DMFA$

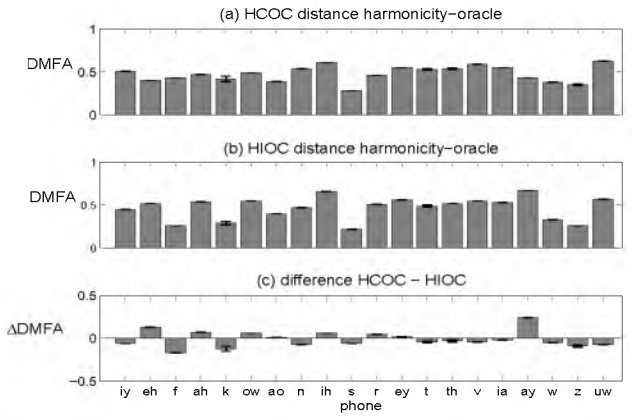

Figure 4. Barchart of the difference *DMFA* between the ***MFA***'s of oracle and harmonicity masks for the (a) *HCOC* subset, (b) the *HIOC* subset, and (c) the difference $\Delta DMFA$ between the two subsets. The distance function took all frequency bands into account with equal weight, using Eq. 5.

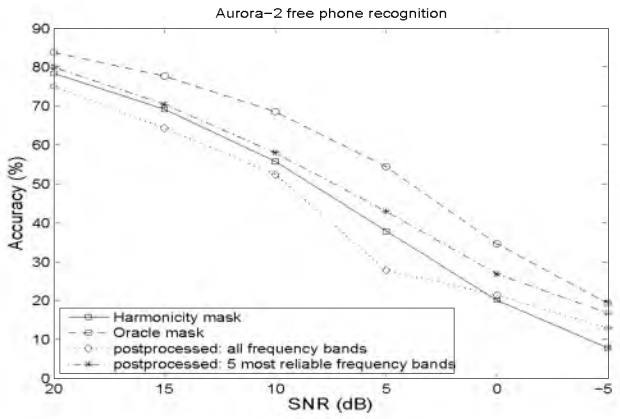

Figure 5. Phone recognition results. Recognition results were obtained using the oracle mask, the harmonicity mask, the post-processed harmonicity mask using all frequency bands and the post-processed harmonicity mask using the 5 frequency bands estimated to be crucial for recognition.

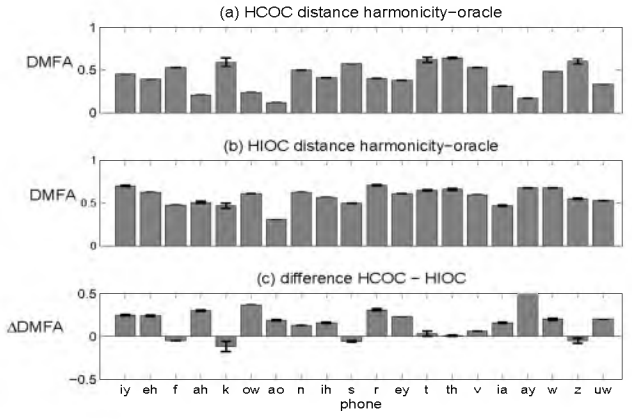

Figure 6. Barchart of the difference *DMFA* between the ***MFA***'s of oracle and harmonicity masks for the (a) *HCOC* subset, (b) the *HIOC* subset, and (c) the difference $\Delta DMFA$ between the two subsets. The distance function only considered a set of frequencies estimated to be the most important for recognition, using Eq. 6.

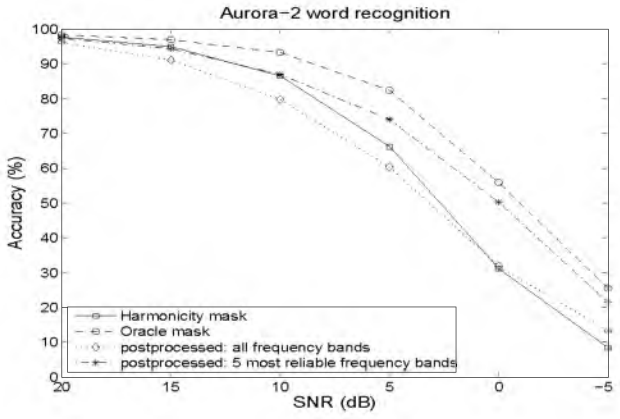

Figure 7. Word recognition results. Recognition results were obtained using the oracle mask, the harmonicity mask, the post-processed harmonicity mask using all frequency bands and the post-processed harmonicity mask using the 5 frequency bands estimated to be crucial for recognition.

Fig. 7 shows recognition results for a word recognizer. Both figures show the approach in which all frequency bands of every phone class were post-processed, as well as the approach in which only the estimated 5 most crucial frequency bands were post-processed. As a reference the accuracies obtained with the original harmonicity and oracle mask are shown.

4.3 Discussion

Recognition results shown in Fig. 5 show once again that a sound-class dependent selection of frequency bands is critical for a correct recognition: Only the post-processing approach using the 5 most reliable frequency bands from the oracle *MFA* leads to an estimated mask with an increased phone recognition accuracy. Allowing *all* frequency bands to change randomly results in a diminished phone recognition accuracy. That an increase in phone recognition accuracy, due to changing mask values, also results in an increase in word recognition accuracy can be observed in Fig. 7.

5 General Discussion

The analysis in Section 3 showed that we can relate the difference (distance) between statistical representations of harmonicity and oracle masks to different accuracies in phone labeling, provided we use a sound-class dependent statistic. The experiment in Section 4 showed that a reduction of this distance leads to increased phone recognition accuracy (Fig. 5), and that this increased recognition accuracy transfers to word recognition (Fig. 7). A statistical analysis of the differences between oracle and harmonicity mask is therefore a valuable tool to analyze and improve techniques for estimating masks.

The failure to predict recognition accuracy by direct comparison of masks as outlined in [8] is due to the non-uniform importance for decoding in time and frequency of the individual mask values. Our proposed statistic, the mask frame average (*MFA*), deals with non-uniform time importance by averaging over homogeneous sound classes. The non-uniform frequency importance is dealt with by selecting sets of frequency bands for every sound class.

The presented analysis technique assumes a cascaded recognition system in which mask estimation is separate from decoding. It is therefore easily adopted for use with other recognition engines, since it treats the decoder as a black box, using only the input mask and the recognition results obtained with this mask. This also ensures that the proposed analysis procedure can also easily be applied to other mask estimation techniques.

The success with which the procedure will lead to a correct analysis of other estimation techniques depends to some extent on the distribution of the mask values over time. Since the *MFA*'s are computed over all time frames in a set, the measure cannot account for local dynamics. This is what actually might have happened with the voiceless plosives, where the spectral envelope differs greatly between the closure and burst. The issue could, however, be addressed by using smaller time intervals such as individual HMM-phone-states to define more homogeneous subsets.

Our analysis tool can be used to give guidance in improving techniques for estimating masks. For example, when it appears that a certain sound class has a consistent distance to the oracle mask, one can take steps to tune the estimation technique to compensate for this. Without this analysis tool, one can only observe the overall increase or decrease in accuracy, without knowing which parts (and therefore which parameters) need to be changed in order to increase recognition accuracy. The analysis framework presented in this paper can also help understanding the results of mask mask estimation techniques in more detail. For example, if a certain estimation technique does not perform better on average for an entire database, but does perform significantly better for certain sound classes, our analysis procedure allows discovering these particular improvements.

The positive results obtained with the distance measure proposed in Eq. 6 shows the need for a sound class dependent mask estimation technique. The experiment presented in section 4 can be considered an approximation of a sound-class dependent mask estimation technique using oracle knowledge of the sound class (by selecting a suitable *MFA*) and general knowledge of the corrupting noise (by using a different *MFA* for every noise type and SNR). The success of this method shows that for proper mask estimation, estimations of both the sound class and some general characteristics of the corrupting noise are critical. The need for sound-class dependent mask estimation makes a cascaded recognition approach infeasible. Our results therefore show that an integrated speech recognition approach, in which mask estimation is combined with decoding, is not only *desireable* but even *crucial*. One such an approach would be to estimate a different mask for every possible sound class and maximizing the joint likelihood of mask and sound class over the utterance. This is somewhat similar to the procedure employed by the speech fragment decoder [10] which selects a mask during decoding by maximizing the likelihood of a collection of mask fragments.

6 Conclusions and future work

We have built an analysis framework which enables us to understand why different spectrographic masks yield different recognition accuracies. We used this framework to study the relation between different masks and the accuracy in a free phone labeling and a digit recognition task. We introduced two distance measures working on the average of coherent sets of mask frames (*MFA*'s). The first includes all frequency bands with equal weight, while the second is limited to a select set of frequency bands. We constructed sets by defining classes of feature vectors with

similar spectral envelopes using a phone based reference transcription. We compared (per frequency band) the statistical differences between oracle and harmonicity masks to the corresponding difference in phone labeling accuracy and explored the effects of decreasing the differences between the two masks on the accuracy of both word and phone recognition.

We conclude that we can relate the difference in recognition accuracy between a harmonicity mask and an oracle mask to the difference between their *MFA*'s by using a sound-class dependent distance function, based on a select set of frequency bands. Our experiments show that it is sufficient to select frequency bands by using frequency bands with the highest a-priori likelihood of being reliable. We can also conclude that biasing an estimated mask toward the statistical properties of the oracle mask results in increased recognition accuracy.

The novel analysis framework can be used as a tool to analyze details of an estimated mask technique and give guidance in tuning the mask estimation. The tool can also be used to analyze the performance of estimated masks in more detail than just the overall increase in recognition accuracy.

The success of the distance measure based on a different subset of frequency bands for every sound class shows that a sound-class dependent mask estimation technique is necessary. We conclude therefore that decoder-based mask estimation technique is not only *desireable* but even *crucial* for improved recognition results.

Ongoing research includes refining the analysis procedure and improving our distance measure. We want to explore to what extent these techniques can be used to predict recognition accuracy for a given estimated mask since this will shorten the development cycle considerably. Furthermore, research is carried out to incorporate sound class estimations in the mask estimation technique as a first step toward a speech decoder in which the mask estimation is an integral part of the decoding process.

Acknowledgments

The research of Jort Gemmeke was carried out in the MIDAS project, granted under the Dutch-Flemish STEVIN program. The project partners are the universities of Leuven, Nijmegen and the company Nuance. The research of Louis ten Bosch is funded in part by the FP6 project ACORNS. We thank Lou Boves for useful discussions.

References

[1] B. Raj, R. Singh, and R. Stern, "Inference of missing spectrographic features for robust automatic speech recognition," in *Proceedings International Conference on Spoken Language Processing*, 1998, pp. 1491–1494.

[2] M. Cooke, P. Green, L. Josifovksi, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, pp. 267–285, 2001.

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

[4] B. Raj, *Reconstruction of incomplete spectrograms for robust speech recognition*, Ph.D. thesis, Camegie Mellon University, 2000.

[5] B. Raj, M. Seltzer, and R. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, pp. 275–296, 2004.

[6] A. Vizinho, P. Green, M. Cooke, and L. Josifovski, "Missing data theory, spectral subtraction and signal-to-noise estimation for robust asr: An integrated study," in *Proceedings of Eurospeech*, 1999, pp. 2407–2410.

[7] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proceedings of IEEE ICASSP*, 2004, vol. 1, pp. 213–216.

[8] 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.

[9] W. Kim and R. M. Stern, "Band-independent mask estimation for missing-feature reconstruction in the presence of unknown background noise," in *Proceedings of IEEE ICASSP*, 2006.

[10] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, vol. 45, pp. 5–25, 2005.

[11] Christophe Cerisara, Sébastien Demange, and Jean-Paul Haton, "On noise masking for automatic missing data speech recognition: A survey and discussion," *Comput. Speech Lang.*, vol. 21, no. 3, pp. 443–457, 2007.

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

[13] ESAT-PSI Speech Group website, http://www.esat.kuleuven.be/psi/spraak/

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