HANDLING TIME-DERIVATIVE FEATURES IN A MISSING DATA FRAMEWORK FOR ROBUST AUTOMATIC SPEECH RECOGNITION

Hugo Van hamme

Katholieke Universiteit Leuven – Dept. ESAT
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

ABSTRACT

We present a novel approach to handling dynamic (time derivative or delta) features for automatic speech recognition using a HMM/GMM-architecture and based on missing data techniques for noise robustness. The static and the dynamic features are imputed in the observations based on an acoustic model expressed in a domain that is a linear transform of the log-spectra and taking bounds into account. The reliability masks of the dynamic features are ternary. We describe a method for computing oracle masks for dynamic features. We also propose a simple method to derive dynamic masks from the reliability mask of the static features. We find that using bounds in the imputation is advantageous, both for oracle masks and for masks derived from the noisy observations.

1. INTRODUCTION

Missing data techniques (MDT) can be applied to alleviate the lack of robustness of speech recognizers to noise. In this approach, spectrographic areas of a noise-contaminated speech signal are considered as unreliable if they are dominated by noise and reliable if the speech signal dominates over the noise signal. The data structure representing this reliability information is called a mask and is expressed with the same time and frequency resolution as the spectrogram. Reliable information is used as such when speech hypotheses are evaluated, but unreliable spectrographic information is either marginalized out or is reconstructed and imputed in the speech spectrum. The mathematical formulation of marginalization and imputation is relatively simple [1] when the speech model is expressed as a HMM with Gaussian mixtures with diagonal covariance matrices in the spectral domain (or any non-linear compression thereof). Every unreliable spectral component can then be marginalized out or can be reconstructed independently of observations and models at other frequencies or other times. In [2], we presented an imputation approach using HMMs with Gaussian mixtures using diagonal covariance matrices in a domain that is a linear transformation of the (log-)spectra. An example of such a representation would be the familiar cepstral coefficients, which we will use in the sequel to develop the ideas. Because the reliability mask is expressed in the spectral domain, but the model is expressed in the cepstral domain, the reconstruction of the missing spectrographic information cannot be solved per frequency bin as before, but requires the solution of a constrained minimization problem involving the data at all frequencies [2]. Because log-spectra and cepstra are related by a linear transform, the problem remains tractable, though significantly more computational effort is involved when compared to a plain vanilla HMM system.

When derivative (dynamic or delta) features are used, the formulation becomes more complex. In addition to a linear transform that mixes over frequency, a linear transform that mixes over time is also introduced. Indeed, the derivatives are computed using a finite impulse response (FIR) filter involving a window of $2L+1$ successive filter bank output values. In [2], all unreliable spectrographic data were reconstructed over this analysis window by likelihood maximization using model involving static and dynamic features. This approach can be criticized for its complexity as well as for its estimation of the missing data based on a model that may be weak for frames with small coefficients of the FIR filters, since they affect the likelihood only marginally. This ill-conditioning could be removed by exploiting the full covariance of the spectral features at different times and frequencies as developed in [3]. In the present work however, a different approach to the MDT-based imputation is proposed. Instead of reconstructing all static features in a window of length $2L+1$, the dynamic features of the central frame are imputed directly. This requires a mask for the delta features. In [5] and [6] so-called *strict* masks are used, i.e. the delta feature is unreliable if any of its contributing static values is unreliable. However, we could not show accuracy improvements with strict masks over a baseline in which the deltas were simply uncompensated.

This paper is organized as follows. In section 2, the basics of missing data based ASR with HMMs are revised. Section 3 describes how dynamic features can be handled in a missing data setting while the speech representation of our choice will be specified in section 4. Section 5

addresses the problem of generating reliability masks for the dynamic features. In section 6 we show that the proposed approach is successful on a small-vocabulary recognition task.

2. MISSING DATA TECHNIQUES FOR ASR

Let $\mathbf{s}_t$, $\mathbf{n}_t$ and $\mathbf{y}_t$ denote the vector of D filter bank outputs at time (signal analysis frame number) t for the clean speech, the noise and the noisy signal respectively. For ease of notation, we will assume throughout this paper that $\mathbf{s}_t$, $\mathbf{n}_t$ and $\mathbf{y}_t$ are log-powers, though other compression functions can be used. Because the noise is additive, the inequality

$$\mathbf{s}_t \le \mathbf{y}_t \tag{1}$$

holds. The reliable components of $\mathbf{s}_t$ are approximated by their counterparts in $\mathbf{y}_t$. In the Gaussian-based imputation approach, the unreliable components of $\mathbf{s}_t$ are estimated as those values that maximize the likelihood of the Gaussian subject to the constraint (1). Hence, for a Gaussian characterized by a mean vector $\breve{\mu}$ and a precision matrix $\breve{\mathbf{P}}$ in the (log-)spectral domain, we need to minimize:

$$\frac{1}{2}(\mathbf{s}_t - \breve{\mu})' \breve{\mathbf{P}} (\mathbf{s}_t - \breve{\mu}) - \frac{1}{2}\log\left(\left|\breve{\mathbf{P}}\right|\right) \tag{2}$$

with respect to the unreliable components of $\mathbf{s}_t$ and subject to (1). In our approach, $\breve{\mathbf{P}}$ is a matrix of full rank containing some structure (see section 4). In case the Gaussians are estimated in the cepstral domain with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, $\breve{\mu} = \mathbf{C}'\mu$ and $\breve{\mathbf{P}} = \mathbf{C}'\Sigma^{-1}\mathbf{C}$ where $\mathbf{C}$ is the orthonormal DCT matrix. By substitution of $\mathbf{x} = \mathbf{y}_t - \mathbf{s}_t$, this minimization is cast as a non-negative least squares (NNLSQ) problem with at most D unknown [2], i.e. the minimization of a quadratic subject to a positivity constraint. We have shown previously that in practice, a few gradient descent iterations suffice to find the solution, leading to a high but often feasible computational complexity. The minimizer of (2) is then imputed in the observation vector to replace the unreliable spectral components.

In practice, $\mathbf{C}$ is often chosen to be non-square, i.e. the number of cepstral coefficients is less than the number D of filters in the filter bank. To have a unique solution of the NNLSQ problem, $\breve{\mathbf{P}}$ needs to be of full rank, which can be achieved by regularizing the problem, e.g. by adding a positive definite diagonal matrix to it.

3. DYNAMIC FEATURES

In the previous section, we described an imputation method for the static spectra. It is, however, very common to augment the feature vector with its first and second order derivatives. In an earlier approach [2], we coped with these delta features in an MDT setting by considering the window of $2L+1$ frames of static features that contribute to the deltas and imputing the missing data over this window. If $\mathbf{b}_{\text{static}}$, $\mathbf{b}_{\text{vel}}$ and $\mathbf{b}_{\text{acc}}$ denote the vectors of $2L+1$ coefficients used to compute the static and dynamic features ($\mathbf{b}_{\text{static}}$ will be all-zeros except for a 1 in the $L+1$-th position), $\otimes$ is the Kronecker product, we can write the augmented cepstral feature vector as

$$\tilde{\mathbf{y}} = \left(\begin{bmatrix} \mathbf{b}_{static} \\ \mathbf{b}_{vel} \\ \mathbf{b}_{acc} \end{bmatrix} \otimes \mathbf{C} \right) \left[\mathbf{y}'_{t-L} \cdots \mathbf{y}'_t \cdots \mathbf{y}'_{t+L} \right]' \tag{3}$$

For every t, all unreliable components of $\left[\mathbf{y}'_{t-L} \cdots \mathbf{y}'_t \cdots \mathbf{y}'_{t+L}\right]'$ can be estimated based on the Gaussian model of $\tilde{\mathbf{y}}$ and the inequalities (1). The resulting non-negative least squares problem has up to $(2L+1)D$ unknown, which makes the high computational load of the approach hard to justify given the modest performance gain relative to using uncompensated deltas. Worse, the estimation of the missing data is an ill-posed problem for those spectral values that contribute little to the deltas.

In the present approach, we define a reliability mask for the delta features themselves and impute the derivative features directly and separately for each feature stream. The resulting problems are well-posed and require significantly less computational effort, but require reliability masks to be made for the deltas (see section 5). Moreover, the mathematical formulation of section 2 needs to be extended.

When static spectral features are unreliable due to noise corruption, this means that the clean value deviates from the noisy observation. With static spectra, the latter is always greater than the former (equation (1)). Similarly, we can define unreliable dynamic spectra as those that deviate from their noisy observations. However, noise corruption can now result in an observation that is either larger or smaller than the clean value. Hence, a reliability detector for dynamic features will need to generate a ternary output (∂ denotes either the velocity or acceleration operator)

$$\begin{aligned} &1 \text{ means unreliable and } && \partial\mathbf{s}_t \le \partial\mathbf{y}_t \\ &0 \text{ means reliable hence } && \partial\mathbf{s}_t = \partial\mathbf{y}_t \\ &\text{-1 means unreliable and } && \partial\mathbf{s}_t \ge \partial\mathbf{y}_t \end{aligned} \tag{4}$$

Since we are using Gaussian mixtures to model the HMM state emission probabilities, the maximum likelihood estimate for the dynamic features is found by minimizing a cost function of the form (2), but now subject to constraints (4). Like for the static features, this problem can be cast as a non-negative least squares problem with at most D unknown.

4. PROSPECT FEATURES AND MDT

In order to reduce the computational load for solving the NNLSQ problems, we abandon the cepstral representation of speech. We replace the DCT matrix $\mathbf{C}$ by another linear

transform **D** that has the property of decorrelating the spectral features [7]. Hence, the PROSPECT features are defined as $\mathbf{p}_t = \mathbf{D}\,\mathbf{s}_t$ with

$$\mathbf{D} = \begin{bmatrix} \mathbf{C}_K \\ \mathbf{I}_D - \mathbf{C}_K{}'\mathbf{C}_K \end{bmatrix} \quad (5)$$

where $\mathbf{C}_K$ is a DCT matrix with K orthonormal rows rendering only the first few cepstral coefficients (K = 3 in this paper) and $\mathbf{I}_D$ is a D-by-D identity matrix. Like cepstra, PROSPECT features can be modeled using a GMM with diagonal covariances [7]. Moreover, this choice makes $\breve{\mathbf{P}}$ of full rank and makes computation of the gradient of (2) more efficient.

In previous work, we have applied PROSPECT models to the static features only. The decorrelation properties of the PROSPECT representation hold equally well for the dynamic features, hence **C** in (3) is replaced by **D** defined in (5). The means and covariances in a PROSPECT model can be estimated using the EM algorithm.

5. MASK ESTIMATION FOR DYNAMIC FEATURES

In the previous sections, we have learned how the missing data can be imputed based on a model and a mask. In this work, we will consider 3 types of masks:

- **Oracle masks** that are derived from the knowledge of the clean speech and the noise. For the static features, the masks are obtained by comparing the log-spectra of clean speech and noise:

$$\mathbf{m}_{static,t} = \left(\mathbf{s}_t \le \mathbf{n}_t - \alpha_{static}\right)_{0/1} \quad (6)$$

  where $(\ )_{0/1}$ equals 1 (0) when the logical expression inside the brackets holds (does not hold) and α_{mask} is a constant. When the cross products of speech and noise are neglected in the Fourier transform of the noisy speech, (6) can be shown to be equivalent to

$$\mathbf{m}_{static,t} = \left(\mathbf{s}_t \le \mathbf{y}_t - \delta_{static}\right)_{0/1} \quad (7)$$

  with $\delta_{static} = \log\left(1 + e^{\alpha_{static}}\right)$. The mask of the dynamic features is defined as:

$$\mathbf{m}_{vel,t} = \left(\Delta\mathbf{s}_t \le \Delta\mathbf{y}_t - \delta_{vel}\right)_{0/1} - \left(\Delta\mathbf{s}_t \ge \Delta\mathbf{y}_t + \delta_{vel}\right)_{0/1} \quad (8)$$

$$\mathbf{m}_{acc,t} = \left(\Delta\Delta\mathbf{s}_t \le \Delta\Delta\mathbf{y}_t - \delta_{acc}\right)_{0/1} - \left(\Delta\Delta\mathbf{s}_t \ge \Delta\Delta\mathbf{y}_t + \delta_{acc}\right)_{0/1} \quad (9)$$

  Hence, the noisy dynamic spectral features are considered reliable if they deviate less than δ from the clean speech values.

- **Derived oracle masks** that are constructed by applying the delta operator of the oracle mask for the static features. Now we set:

$$\mathbf{m}_{vel,t} = sign\left(\left[\mathbf{m}_{static,t-L} \ldots \mathbf{m}_{static,t} \ldots \mathbf{m}_{static,t+L}\right]\mathbf{b}'_{vel}\right) \quad (10)$$

$$\mathbf{m}_{acc,t} = sign\left(\left[\mathbf{m}_{static,t-L} \ldots \mathbf{m}_{static,t} \ldots \mathbf{m}_{static,t+L}\right]\mathbf{b}'_{acc}\right) \quad (11)$$

  where $sign$(x) is 1, 0 or –1 if x is positive, zero or negative. This choice is motivated as follows. The derivative spectra are a linear combination of the static spectra. The static masks flag the fact that the noisy contribution to this linear combination is less than the clean value. With positive (negative) **b**-weights, the noisy derivative will be less (greater) than the clean value. Hence, we can consider (10) and (11) as a weighted voting mechanism. The static features are considered reliable when there are equal votes for over and underestimation due to the noise corruption. This happens in particular when all features are reliable, but also when all spectral features are unreliable.

- **Derived real masks** that are constructed for the dynamic features by applying the delta operator of a real mask for the static features. Now $\mathbf{m}_{static,t}$ is estimated from the noisy data using harmonicity and SNR information [4]. The dynamic masks are constructed with (10) and (11).

6. EXPERIMENTS

The above approach is evaluated on the AURORA-2 continuous digit recognition task. Since channel mismatch is beyond the scope of this paper and no prior knowledge about the noise is exploited, we limit the evaluation to test set A. The recognizer is configured with 16 HMM states per digit and 20 Gaussians per state. The optional inter-word silence is modeled by 1 or 3 states with 36 Gaussians per state, while leading and trailing silence have 3 states. The total number of Gaussians is 3628. The front-end of the MDT system is the ETSI STQ WI-007 standard, using 23-channel MEL-spaced filter bank and no (cepstral) mean normalization. The filter bank outputs are transformed to PROSPECT features with K = 3. Velocity and acceleration features are computed using the HTK default regression formulae. First, a reference model was trained using the standard AURORA training script and using WI-007 MFCCs with their first and second order deltas. The Gaussian means and diagonal covariances in the PROSPECT domain are obtained by "single-pass retraining", i.e. forced alignment using noise-free cepstral features while the accumulants of the EM training are computed for the noise-free PROSPECT features.

The mean accuracy over the four noise types of test set A is presented in figure 1. Curve (a) shows the baseline without noise compensation and using the PROSPECT transformation on static and dynamic features. On clean data, the error rate is 0.4%, which is even better than the cepstral model's performance. Hence static as well as dynamic features are well-modeled with Gaussians with a diagonal covariance in the PROSPECT domain. Curve (b) shows the accuracy when the oracle masks (6) with α_{static}=3dB are used for imputation of the static features while the dynamic features are uncompensated. When compared to the results in [7] which used the PROSPECT transform only on the

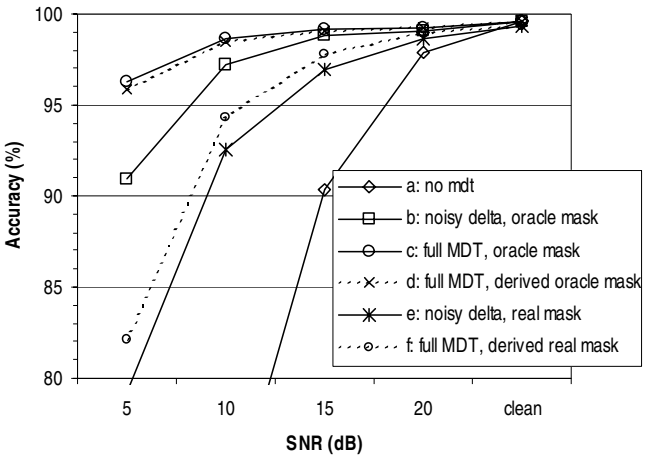

Figure 1: recognition accuracy using the PROSPECT model for static, velocity and acceleration features and (a) no noise compensation (b) MDT with oracle masks on the statics only (c) oracle masks and MDT on all features (d) derived oracle masks and MDT on all features (e) MDT with real masks on the statics only (f) MDT with derived real masks.

static features and cepstra on the dynamic features, we see almost equal performance, which extends our previous conclusion to low SNR.

Next, we try to handle the delta features with missing data techniques. By imputing the dynamic features using oracle masks (8) and (9) with δ_{vel}=2dB and δ_{acc}=0.5dB and bounds (4), a superior accuracy is obtained (curve c). This performance degrades only slightly when the derived masks (10) and (11) are used instead (curve d) and is definitely better than no compensation for the deltas (curve b). Strict masks (see section 1) in conjunction with imputation without bounds resulted in practically equal performance as curve (b). Though results with uncompensated deltas were not reported in [5], they found small differences with exploiting bounds on deltas, while we find a significant improvement. We attribute this contradiction to differences in acoustic models, domain and MDT method (marginalization versus imputation).

In a last series of experiments, we replace the static oracle mask by one computed from the noisy data as described in [4]. When the delta features are left uncompensated, we obtain curve (e). Using strict masks with imputation without bounds we obtained worse results than this reference (e.g. 3 % absolute at SNR of 10dB). The difference with oracle masks is that these real strict masks become sparse so the delta features are almost always ignored. By using derived real masks (which are less sparse) and exploiting the bounds, we obtain the curve (f), showing an improvement attaining the performance reported in [4], but now in the absence of additional noise reduction methods. Unlike the conclusion from [5], we find that bounds for the deltas do help on real masks.

7. DISCUSSION AND CONCLUSIONS

We have proposed a data imputation method for handling the streams of dynamic features in speech recognizers based on missing data techniques, where the observations are used as upper or lower bounds as indicated by the reliability mask. The missing data were imputed in the static as well as in the dynamic features based on a HMM for speech using Gaussian mixture emission models. In our missing data approach, these Gaussians are not constrained to be diagonal in the spectral domain, but are diagonal in any linear transform of the spectral domain. For reasons of computational efficiency, we opted for the PROSPECT transform, which was shown to be a valid replacement of the cepstral transform. We extended the application of this PROSPECT transformation to the delta features and showed experimentally that this leads to comparable performance. Then we showed the validity of the imputation method of the deltas based on oracle reliability masks and proposed a simple method to compute a reliability mask for the dynamic features from a static mask. We find that bounded imputation of dynamic features improves accuracy for both oracle and real masks.

8. REFERENCES

[1] Cooke, M., Green, Ph., Josifovski, L., Vizinho, A. "Robust automatic speech recognition with missing and unreliable acoustic data". *Speech Communication 34* (*2001*), pp. 267-285

[2] Van hamme, H., "Robust Speech Recognition Using Missing Feature Theory in the Cepstral or LDA Domain," *Proc. Eurospeech*, Geneva, Sept. 2003, pp. 3089-3092.

[3] Raj, B., Seltzer, M.L., Stern, R.M. "Reconstruction of missing features for robust speech recognition". *Speech Communication 443 (2004),* pp. 275-296.

[4] Van hamme, H, "Robust speech Recognition using cepstral domain missing data techniques and noisy Masks," In *Proc. ICASSP*, pp. 213-216, Montreal, May 2004.

[5] Josifovski, L. "Robust automatic speech recognition with missing and unreliable data", PhD thesis, University of Sheffield, 2002

[6] Yamamoto, S, Valin, J.-M., Nakadai, K., Rouat, J., Michaud, F. Ogata, T., Okuno, H. "Enhanced robot speech recognition based on microphone array source separation and missing feature theory," In *Proc. Interspeech*, Lisbon, Portugal, September 2005

[7] Van hamme, H, **"**PROSPECT Features and their Application to Missing Data Techniques for Robust Speech Recognition," In *Proc. ICSLP,* pp. 101-104, Jeju Island, Korea, October 2004.