Robust Speech Recognition Using Missing Feature Theory in the Cepstral or LDA Domain

Hugo Van hamme

Katholieke Universiteit Leuven – Dept. ESAT
Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium
Hugo.Vanhamme@esat.kuleuven.ac.be

Abstract

When applying Missing Feature Theory to noise robust speech recognition, spectral features are labeled as either reliable or unreliable in the time-frequency plane. The acoustic model evaluation of the unreliable features is modified to express that their clean values are unknown or confined within bounds. Classically, MFT requires an assumption of statistical independence in the spectral domain, which deteriorates the accuracy on clean speech. In this paper, MFT is expressed in any domain that is a linear transform of (log-)spectra, for example for cepstra and their time-derivatives. The acoustic model evaluation is recast as a non-negative least squares problem. Approximate solutions are proposed and the success of the method is shown through experiments on the AURORA-2 database.

1. Introduction

The speech units of an acoustic model of a speech recognizer are trained on many examples of their occurrence. This learning approach has poor generalization properties when there is a mismatch between training and testing conditions and results in poor robustness. One source of mismatch is caused by additive noise to the speech signal. While other sources of mismatch such as dialects, accents, Lombard effects, … are difficult to understand and to model, the effect of noise addition on the extracted features is a tractable calculation. Hence it is somehow a shame to use mixed style training techniques to account for the impact of noise.

In this paper, the mismatch between clean training examples and noisy testing features will be reduced by applying Missing Feature Theory (MFT) [1]. A missing feature detector will identify at any given time which components of a feature vector are corrupted by the noise. Those so-called *unreliable* features will be treated differently in the acoustic model evaluation, to express that their values have different statistics than those of clean speech. The reliable components on the other hand are treated as if they have been computed in the absence of noise.

The missing feature models holds fairly well in the spectral domain. When examining noisy spectrograms, one can observe that the areas of low speech energy in the time-frequency plane are replaced by the noise spectra, while the high energy zones, such as the formants of voiced speech are only affected for high noise levels. A major problem with the application of MFT to current state-of-the-art speech recognizers is that they use cepstral or other features which are obtained from a spectral representation by linear transformations. While the noise contamination is localized in the time-frequency plane, all transformed features are usually corrupted if one spectral value is unreliable. Hence, most authors have applied MFT in the spectral domain, using Gaussian mixtures with diagonal covariance. This model is known to be inferior with respect to for example a model in the cepstral domain. The goal of this paper is to formulate the MFT approach for the features and an acoustic modeling that is used by most mainstream speech recognizers.

MFT is applied in several variants. One MFT approach for dealing with unreliable features called *marginalization* is to disregard them, which is motivated by the reasoning that missing features carry no information. In [2], this method is applied in the cepstral domain. The approach that will be used in this paper on the other hand will infer the clean speech values of the missing features by taking an upper bound on these values into account. Indeed, when noise is added to the signal, the unknown clean speech energy at a particular place in the time-frequency plane will be lower than the observed noisy energy. Notice that this is an approximation since speech and noise might also cancel each other.

The paper is organized as follows: section 2 describes the setting to which this paper applies, section 3 describes a few preferred approaches of MFT to noise robust speech recognition and recasts Gaussian evaluation using MFT as a non-negative least squares problem. Experimental evaluations are provided in section 4. Final considerations, suggestions for further work and conclusions are provided in sections 5 and 6.

2. The speech recognizer

This study focuses on noise that is additive to the speech signal. Unknown filtering will not be considered in this paper. The speech recognizer is assumed to have a mainstream HMM-based architecture with Gaussian mixture acoustic models. In the front-end, a low-resolution spectral representation is computed by a filter bank through windowing, framing, FFT, filter bank integration and nonlinear compression. For each frame t, this results in a K-dimensional spectral vector $y(t)$. A typical example would be the computation of a MEL-scaled log-spectrum with a fixed frame rate as it will be used in section 4 of this paper. Subsequently, the vectors from a sliding window of length $2M+1$ centered around t are stacked to form the augmented spectral vector $y'(t)$:

$$y'(t) = \begin{bmatrix} y(t-M) \\ \vdots \\ y(t) \\ \vdots \\ y(t+M) \end{bmatrix} \quad (1)$$

Finally, this augmented spectral vector is transformed linearly to form the feature vector $f(t)$ of dimension D:

$$f(t) = C \, y'(t) \qquad (2)$$

where C is a real-valued D-by-$(2M+1)K$-matrix. In case of cepstral features with time derivatives, C would take the form:

$$C = \begin{bmatrix} b_{static} \\ b_{delta} \\ b_{acc} \end{bmatrix} \otimes C_{DCT} \qquad (3)$$

where C_{DCT} is the (truncated) DCT matrix and $\otimes$ denotes the Kronecker product. The $2M+1$ dimensional row vectors b_{static}, b_{delta} and b_{acc} are the FIR filter coefficients to compute static, delta and acceleration features, i.e. b_{static} is all zeros except for the central entry and b_{delta} could be chosen as one of the well-known regression formulas. In case of LDA-features, C could be a full matrix.

The acoustic model assumes a Gaussian mixture for the state probability density functions in the f-space:

$$P\big(f(t)\,|\,q\big) = \sum_i w_{iq} N\big(f(t); \mu_{iq}, \Sigma_{iq}\big) \qquad (4)$$

where q denotes the HMM state, i is the mixture index and $N(x \, ; \, \mu, \Sigma)$ denotes a D-dimensional multivariate Gaussian density function at x with mean μ and covariance Σ (usually diagonal). Instead of using (4) directly, the equivalent HMM formulation will be used where every state q is decomposed into a set of fully connected expanded states (i,q), each with density

$$P\big(f(t)\,|\,i,q\big) = N\big(f(t); \mu_{iq}, \Sigma_{iq}\big) \qquad (5)$$

The transition probabilities are adjusted appropriately with the mixture weights. This step is made for theoretical convenience; in practice, the expansion of the HMM does not need an explicit implementation.

3. Problem formulation

The following notations will be used for the spectral K-vectors: $y(t)$ for the noisy speech, $s(t)$ for the clean speech and $n(t)$ for the noise. Similarly, a prime will be added to denote the corresponding $(2M+1)K$-vectors of stacked frames. The subscripts u and r will be used to denote the unreliable and reliable components of a vector respectively.

Because of the noise addition assumption, the noisy spectral frame can be expressed as a function of the clean speech spectrum, the noise spectrum and their phase differences. This yields the approximation that in any of the K filter bank channels, either the speech or the noise dominates the output:

$$y(t) \approx \max\big(s(t), n(t)\big) \qquad (6)$$

where the max-operator works component-wise over its 2 arguments. Equation (6) also leads to the approximate constraint:

$$s_u(t) \leq y_u(t) \qquad (7)$$

where the inequality holds component-wise. This inequality will be used as a constraint for estimating the missing clean speech spectral values.

3.1. MFT in the spectral domain

If $s(t)$ is modeled by a sum of Gaussian mixtures with diagonal covariance, then the "data imputation using bounds"

method described in section 3.6 of [1] can be applied. For every Gaussian, any missing feature component is replaced by the corresponding component of the Gaussian mean μ_{iq} or the bound $y(t)$ in equation (7), whichever is the smaller one. This value maximizes the expanded state likelihood (5) under constraint (7) and is hence the (i,q)-conditional maximum likelihood estimator of the clean speech given the noise. In [1], these estimates are combined with the posterior mixture probabilities to from a q-conditional estimate. However, on an AURORA-2 evaluation (not reported on here), it was observed that better accuracy is obtained if each Gaussian is evaluated in its (i,q)-conditional estimate. Hence, in this paper, clean speech estimates will also be conditioned on the expanded state or equivalently on the mixture component.

3.2. MFT in the cepstral domain

As pointed out above, this section is equally valid for any linear transformation (2) instead of the cepstral one.

An expanded-state-conditional maximum likelihood estimate of the missing components of the clean speech in $s'(t)$ can be obtained by maximizing (5) subject to the constraints (7). These constraints need to be expressed for every frame in the window $[t\text{-}M, \ldots, t, \ldots, t\text{+}M]$. By maximizing the logarithm of (5) and neglecting the constant terms, the problem becomes the minimization over s' of the cost function

$$\big(C \, s' - \mu_{iq}\big)^t \Sigma_{iq}^{-1} \big(C \, s' - \mu_{iq}\big) \qquad (8)$$

subject to

$$s'_u \leq y'_u \text{ and } s'_r = y'_r \qquad (9)$$

For notational convenience, the frame-dependence was dropped. By grouping the columns of C into those that are multiplied by the reliable and unreliable features, yielding the matrices C_r and C_u, the cost function (8) is rewritten as

$$\big(C_u \, s'_u + C_r \, y'_r - \mu_{iq}\big)^t \Sigma_{iq}^{-1} \big(C_u \, s'_u + C_r \, y'_r - \mu_{iq}\big) \quad (10)$$

Finally, by substituting $s'_u = y'_u - x$ and by observing that the quadratic cost is associated with an equivalent least squares problem, the non-negative least squares (NNLS) problem is obtained:

$$\Sigma_{iq}^{-\frac{1}{2}} C_u \, x \approx \Sigma_{iq}^{-\frac{1}{2}} \big(C \, y' - \mu_{iq}\big) \text{ with } x \geq 0 \qquad (11)$$

In most cases, C_u has more columns than rows and (11) has an infinite number of solutions. Therefore, the NNLS problem will be augmented with additional equations

$$\begin{bmatrix} \Sigma_{iq}^{-\frac{1}{2}} C_u \\ \lambda \, H \end{bmatrix} x \approx \begin{bmatrix} \Sigma_{iq}^{-\frac{1}{2}} \big(C \, y' - \mu_{iq}\big) \\ \lambda \, h \end{bmatrix} \text{ with } x \geq 0 \quad (12)$$

where λ is a fixed weight.

For a generic choice of λ, h and H, the solution of the augmented problem (12) will defer from the original solution, but the impact can be controlled through λ. To avoid underdetermination, the augmented matrix on the left hand side of (12) should be of full rank. The complexity of the augmented problem will be kept low by choosing a sparse structure for H. In practice, h and H will be chosen to express a property of s', such as the correlation over time due to the framing overlap, the correlation over frequency due to the filter bank overlap, or an approximate covariance structure of s'.

3.3. Solving the NNLS problem

Solving (12) for x is a standard mathematical problem that requires iteration. Because of the sparse structure, optimized techniques can be applied [3]. The exact solution of the non-negative least squares problem will be obtained using the MATLAB *snnls*-function from [4].

While exploiting the sparse structure of H speeds up the computation significantly, it is still too heavy for practical implementation. A first approximation - denoted by IT0 - is given by solving the sparse NNLS problem $H\,x \approx h$ only. A second approximation, denoted by IT1, is obtained by performing one gradient search step from the initial point IT0. The step size is controlled appropriately such that $x \geq 0$ is not violated.

3.4. Marginalization in the cepstral domain

In marginalization [1], unreliable feature components are removed from the evaluation of a Gaussian. In the cepstral domain, this approach was described in [2], by inserting a diagonal weighting matrix W (ideally with weights 0 or 1) in the Gaussian exponential expression:

$$\left(y - \tilde{\mu}_{iq}\right)^t W^t C_{DCT}^t \Sigma_{iq}^{-1} C_{DCT} W \left(y - \tilde{\mu}_{iq}\right) \tag{13}$$

where the tilde on the state mean indicates that it is computed in y-space (alternatively it can be obtained through an inverse of C_{DCT}). While this approach is simple, like marginalization in the spectral domain, it does not use the bounds (7), which results in inferior accuracy with respect to the current approach.

4. Experiments

In this section, the theory outlined above will be evaluated for the special case of cepstral features, possibly with their derivatives.

The choice for h and H used in this paper will be derived from the second order statistics of the y or y' vectors, i.e. for every Gaussian (i,q) of the acoustic model in the cepstral domain, an *auxiliary* Gaussian will be estimated which is described by the mean μ_a and covariance Σ_a of y or y' for that expanded state. The covariance structure will be approximated by a diagonal one, which leads to a diagonal H matrix:

$$H = \Sigma_a^{-\frac{1}{2}} \text{ and } h = \Sigma_a^{-\frac{1}{2}}\left(y - \mu_a\right) \tag{14}$$

Notice that when applied to y, the solution of $H\,x \approx h$ is identical to the "expanded state conditional data imputation with bounds MFT-method" as explained in section 3.1. The diagonal structure of H allows the independent solution of the least squares problems per dimension.

The experiments are based on an idealized missing feature detector, namely a features are unreliable if and only if $s(t) \leq n(t)$. In a real system, this component needs to be replaced by one that doesn't use instantaneous speech and noise knowledge. The performance of a missing feature detector depends on the assumptions one is willing to make about the noise. Since the goal of this paper is to describe the framework for expressing MFT in the cepstral domain, the impact of these assumptions are factored out by the idealized detector.

4.1. AURORA-2 database

The speech recognition task is to recognize digit strings at a bandwidth of 8 kHz embedded in artificially added noise as defined in the AURORA-2 database [5]. Each recognition experiment at a given SNR and noise type consists of 1001 utterances, comprising a total of over 3000 digits. The experimental verification will be based on test set A for the non-stationary suburban train noise (N1).

4.2. Acoustic models

The front-end of the speech recognizer is the reference implementation Track 1 v2.0 as included in the AURORA distribution. First, baseline models were trained on the clean training data using the AURORA reference script for 12 cepstral coefficients and c_0 (no log-energy), their velocity and their acceleration features. These acoustic models will be referred to as CEP_D_A. All digits are represented by 16 HMM states, connected strictly left-to-right without skips, each with 3 Gaussians, without state tying. The inter-word silence model has a single state and the leading/trailing silence model has 3 HMM states with state transition structure given in [5] and 6 Gaussian mixture components.

Some experiments are performed on static features only. The CEP model is derived from the CEP_D_A model by removing the dynamic parameters from the state densities and performing 5 Baum-Welsh iterations on the static cepstra.

The auxiliary models are obtained by accumulating statistics of the auxiliary features during the last Baum-Welsh iteration of the training process of the clean speech models using the HTK HERest function in "single pass training" mode. In this way, every Gaussian of the CEP_D_A or CEP model gets and associated Gaussian with diagonal covariance whose statistics are computed on the auxiliary feature stream.

The decoder and acoustic model scoring are implemented in locally developed software that uses the same decoding grammar and transition penalties as the HTK reference of the AURORA benchmark.

4.3. Static features

For the sake of simplicity, the method is first examined on static cepstra where $K = 23$, $M = 1$ and $D = 13$. The accuracy obtained with the CEP models serves as the baseline in Fig. 1 (doubly dashed line). While the accuracy under clean conditions is fairly good, the model exhibits hardly any robustness to noise. The marginalization approach of section 3.4 can fix some of the lack of robustness as depicted by the line called MARGIN. The accuracy obtained using the exact NNLS solution as described in section 3.3 and (14) with $\lambda = 0.5$ is shown as the dashed line (NNLS) and exhibits far greater robustness. This performance is well approximated by IT0, which corresponds to constructing the MFT solution in the spectral domain and evaluating the score in the cepstral domain. When one gradient descent iteration is applied, the likelihood scores increase as well as the accuracy as depicted by the curve IT1 in Fig. 1. Strangely, IT1 even outperforms the exact NNLS solution. This observation needs further investigation, but it could be explained by the over-estimation of the Gaussian likelihood at the optimal NNLS solution, a phenomenon reported upon in a separate paper and which applies to a lesser extent to sub-optimal solutions such as IT1.
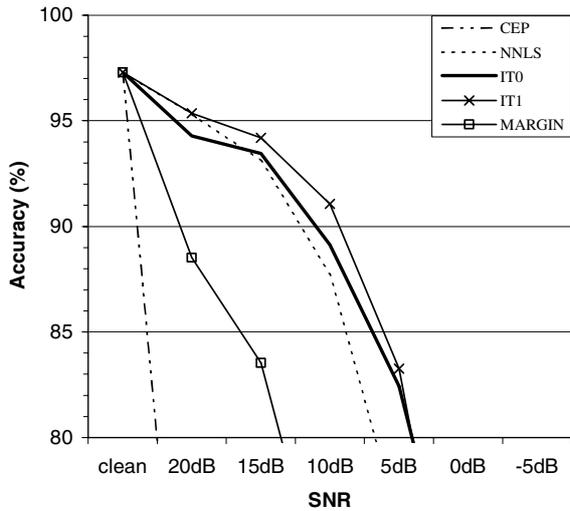
Figure 1: *accuracy on the AURORA test set A (train noise) using static features only.*

Figure 2: *accuracy on the AURORA test set A (train noise) using static and dynamic features.*

4.4. Dynamic features

In the reference front-end of the AURORA-2 task, the delta cepstra are computed on a 7 frame window and the acceleration features are computed on a 11 frame window (frame rate is 100Hz). Hence, the values $K = 23$, $M = 5$ and $D = 39$ are used in this section. While this is longer than is probably optimal for the proposed approach, the standard window length was maintained.

An overview chart of the obtained accuracy is given in Fig. 2. Compared to the results for static cepstra, both the clean accuracy and the noise robustness benefit from the addition of dynamic parameters in the CEP_D_A baseline. The full NNLS solution is too expensive in this case due to the high dimension. The approximate solution IT1 (solid line with circles) improves the robustness at the higher noise levels, but although the actual recognition errors are different from the CEP_D_A baseline, leads to the same accuracy at 20 dB SNR.

Since dynamic features are more robust to noise addition than statics, an alternative and cheaper method is to apply the NNLS method to the static cepstra like in section 4.3 and use the noisy dynamic cepstra for the remaining 26 feature components. The accuracy for this method is given by the solid curve using the IT1 approximation. This method exceeds the performance for the multi-style training approach [5] (dashed line), though it needs to be repeated that the missing feature detector was idealized in these experiments. The poorer performance of the full IT1 method is probably due to the inappropriateness of a diagonal covariance model for *s'*, which spans 11 frames.

5. Discussion and future work

The IT0 and IT1 approximations lead to simple algorithms for solving the NNLS problem. The main computational complexity that is left in the Gaussian likelihood evaluation are multiplications with *C*, e.g. the transformation of the MLE sol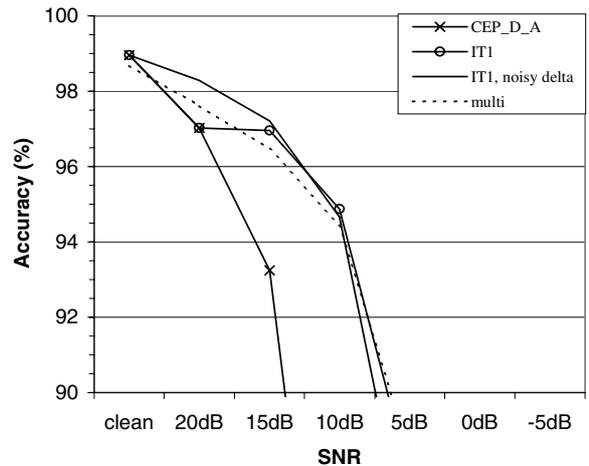ution for *s'* in (8). However, many of these solutions will coincide or can be cached. Moreover, the likelihood of the auxiliary Gaussian is an important cue for a pruning criterion. Future work will focus on different choices for *H* and *h*, on scrutinizing the approximate NNLS solutions, on pruning strategies, on the use of a non-idealized missing feature detector and on more extended test sets.

6. Conclusions

A framework for applying MFT in linear transform domains of spectra was outlined. The formulation leads to a non-negative least squares (NNLS) problem. One approximate solution of practical importance is given by computing the MFT-based "data imputation" estimator in the log-spectral domain and evaluating its score in the transformed domain. One gradient search iteration further improves the accuracy. This approach was verified for both static and static plus dynamic cepstral features on a subset of the AURORA-2 task.

7. References

[1] Cook 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] Häkkinen J., Haverinen H. "On the Use of Missing Feature Theory with Cepstral Features", *Proc. CRAC Workshop, Aalborg, Denmark, September 2, 2001*

[3] Adlers M. *Topics in Sparse Least Squares Problems*, PhD thesis, April 2000, Linköping University, Sweden

[4] http://www.mai.liu.se/~milun/sls/

[5] Hirsch H.-G., Pearce D. "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions". *Proc. ISCA ITRW ASR2000 Workshop, Paris, France, September 18-20, 2000*