STATE BASED IMPUTATION OF MISSING DATA FOR ROBUST SPEECH RECOGNITION AND SPEECH ENHANCEMENT

Ljubomir Josifovski, Martin Cooke, Phil Green and Ascension Vizinho

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello St, Sheffield S1 4DP, UK
{L.Josifovski,M.Cooke,P.Green,A.Vizinho}@dcs.shef.ac.uk

ABSTRACT

Within the context of continuous-density HMM speech recognition in noise, we report on imputation of missing time-frequency regions using emission state probability distributions. Spectral subtraction and local signal–to–noise estimation based criteria are used to separate the present from the missing components. We consider two approaches to the problem of classification with missing data: marginalization and data imputation. A formalism for data imputation based on the probability distributions of individual Hidden Markov model states is presented. We report on recognition experiments comparing state based data imputation to marginalization in the context of connected digit recognition of speech mixed with factory noise at various global signal-to-noise ratios, and wideband restoration of speech. Potential advantages of the approach are that it can be followed by conventional techniques like cepstral features or artificial neural networks for speech recognition.

1. INTRODUCTION

There is evidence that an ability to deal with masked or missing data is inherent to human audition at various levels. Listeners show robust performance with intentional occlusions (both across time and frequency) and with band–restricted speech [12]. Neurophysiological evidence suggests that the neural response to weaker signals is masked and can be considered lost for subsequent auditory processing [7]. Sound sources are also obscured in a mixture, yet, the partial information available to the listeners is often sufficient for recognition due to redundant coding of information in speech [10].

The missing data approach to speech recognition [2, 6] fits well with computational auditory scene analysis (CASA) [1] techniques for sound separation, since it makes no prior assumptions about the number or the stationarity of the sources. However, if additional knowledge is available, it is easily utilised and integrated in this framework.

There are two subproblems in the application of the missing data techniques in robust ASR: identification of the reliable spectro–temporal regions and the need for classification techniques that can deal with incomplete data. We apply spectral subtraction with an additional SNR criterion to the first problem. In the context of an HMM based system, section 2 describes two solutions to the problem of classification with incomplete data based on marginalization of the state emission p.d.f. Section 3 introduces a novel approach that is based on prediction of the missing values from the conditional state p.d.f. Section 4 describes the results of the experiments carried out on the TIDigits corpus with added NOISEX noise, and investigates the feasibility of wide–band speech restoration. A complementary paper in this proceedings describes marginalization in conjunction with SNR estimation [11].

2. CLASSIFICATION WITH MISSING DATA

It is assumed that the input data vector x has been partitioned into a reliable part x_r and an unreliable part x_u in the previous stage of processing. The state emission distributions $f(x|S)$[1] (inferred during training) require the full feature vector $x = (x_r, x_u)$. Instead of computing the likelihood $f(x_r, x_u|S)$ of the full data, the likelihood of the data that is reliable/present can be computed [2, 3, 6, 8]:

$$f(x_r|S) = \int f(x_r, x_u|S) dx_u. \tag{1}$$

In the context of a continuous density HMM system, we can model the state distributions as mixtures of M multivariate Gaussians with diagonal covariance matrices:

$$\begin{aligned} f(x|S) &= \sum_{k=1}^{M} P(k|S) f(x|k, S), \\ &= \sum_{k=1}^{M} P(k|S) f(x_r, x_u|k, S), \end{aligned} \tag{2}$$

where $P(k|S)$ are the mixing coefficients. The features within the Gaussians in the mixture are independent, and thus:

$$f(x_r, x_u|k, S) = f(x_r|k, S) f(x_u|k, S). \tag{3}$$

[1] In this paper, $f(\cdot)$ denotes probability density, while $P(\cdot)$ denotes probability.

Proceeding with the marginalization in (1):

$$\begin{aligned} f(x_r|S) &= \int \sum_{k=1}^{M} P(k|S) f(x_r, x_u|k, S) dx_u, \\ &= \sum_{k=1}^{M} P(k|S) f(x_r|k, S) \underbrace{\int f(x_u|k, S) dx_u}_{1}, \\ &= \sum_{k=1}^{M} P(k|S) f(x_r|k, S). \end{aligned} \quad (4)$$

We refer to this method for computing the likelihood of the incomplete data vector as *marginalization*.

Further, knowledge that the unreliable data is bounded can be utilized to integrate x_u in the range (x_{low}, x_{high}) instead of $(-\infty, +\infty)$:

$$\begin{aligned} f(x_r|S) &= \int_{x_{low}}^{x_{high}} \sum_{k=1}^{M} P(k|S) f(x_r, x_u|k, S) dx_u, \\ &= \sum_{k=1}^{M} P(k|S) f(x_r|k, S) \int_{x_{low}}^{x_{high}} f(x_u|k, S) dx_u. \end{aligned} \quad (5)$$

The integral can be evaluated in the case of Gaussian distributions using the standard error function. We refer to this method for computing the likelihood of the incomplete data vector as *bounded marginalization*.

The marginal distribution of the components of the diagonal Gaussian mixture $f(x_r|k, S)$ is diagonal Gaussian itself and is readily computable.

3. STATE BASED DATA IMPUTATION

Instead of computing the likelihood given the data present, it is also possible to compute the distribution of the unreliable parts of the feature vector using the reliable components and the joint p.d.f. Then, some representative value for the unreliable/missing feature can be chosen using this distribution [5, 9]. Usually it is the mean of the conditional distribution. Once missing features are reconstructed, ASR can proceed as with complete data.

In state based data imputation (SDI), we use HMM state distributions to infer the conditional distribution $f(x_u|x_r, S)$. Again, we assume that state distributions are mixtures of multivariate diagonal Gaussians (2):

$$\begin{aligned} f(x_u|x_r, S) &= \frac{f(x_u, x_r|S)}{f(x_r|S)} = \frac{\sum_{k=1}^{M} P(k|S) f(x_r, x_u|k, S)}{f(x_r|S)}, \\ &= \sum_{k=1}^{M} \frac{f(x_r|k, S) P(k|S)}{f(x_r|S)} f(x_u|k, S), \\ &= \sum_{k=1}^{M} P(k|x_r, S) f(x_u|k, S), \end{aligned} \quad (6)$$

where:

$$P(k|x_r, S) = \frac{P(k|S) f(x_r|k, S)}{f(x_r|S)} = \frac{P(k|S) f(x_r|k, S)}{\sum_{k=1}^{M} P(k|S) f(x_r|k, S)}, \quad (7)$$

can be considered to be the *responsibility* of the components. The form (6) is very convenient, since it too is a mixture of diagonal Gaussians.

The mean of the conditional distribution is:

$$\begin{aligned} E_{x_u|x_r,S}\{x_u\} &= \int f(x_u|x_r, S) x_u dx_u, \\ &= \sum_{k=1}^{M} P(k|x_r, S) \underbrace{\int f(x_u|k, S) x_u dx_u}_{\mu_{u|k,S}}, \\ &= \sum_{k=1}^{M} P(k|x_r, S) \mu_{u|k,S}. \end{aligned} \quad (8)$$

This value is readily computable as the sum of the means of the unreliable features from all components of the mixture, weighted by their responsibilities. The responsibilities (7) are recomputed mixture weights after the reliable data was identified.

Once the conditional mean of each unreliable feature is computed, the values of x_u in the feature vector are replaced by $E_{x_u|x_r,S}\{x_u\}$. In every time frame the imputations of all HMM states are computed. Hence there are as many versions of the frame as there are states. During the emission probability calculation, for each state, only the likelihood of the feature vector with x_u's filled from the p.d.f. of the same state is computed. Therefore, the complexity of the search for the best model/state alignment is of the same order as standard ASR.

4. EXPERIMENTS

4.1. Experimental setup

The TIDigits corpus of digits sequences was used. Acoustic vectors consisted of smooth outputs of from 64–channel auditory filter bank (centre frequencies spaced linearly in ERB-rate from 50 to 8000Hz), computed every 10ms. HTK [13] was used for training, and a local MATLAB decoder for recognition. Twelve models ('1'–'9', 'oh', 'zero' and 'silence') consisting of 8 no–skip, straight–through states with observations modeled with a 10 component diagonal Gaussian mixture were trained on clean speech. Non–stationary factory noise from NOISEX was added (with random start points) at SNRs from +20dB to -5dB to a subset of the TIDigits test set consisting of 240 digit strings used for testing. For bounded marginalization, the bounds were set to 0 and the value of the noisy speech mixture at each time–frequency point.

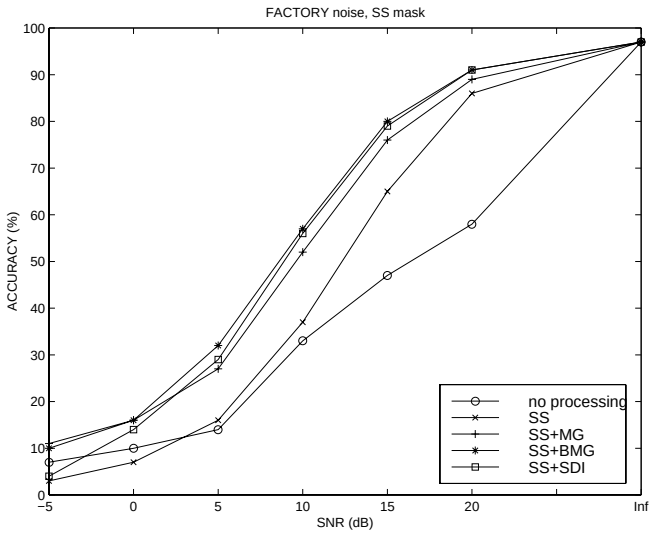

Figure 1: Accuracy of bounded and unbounded marginalization and unbounded data imputation with spectral subtraction criterion (9).

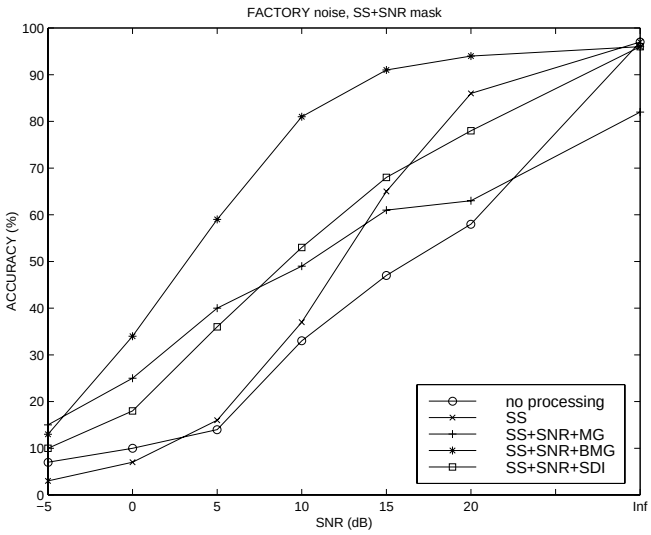

Figure 2: Accuracy of bounded and unbounded marginalization and unbounded data imputation with spectral subtraction (9) and SNR (10) criterias.

4.2. Identifying unreliable data

In the first set of experiments, spectral subtraction (SS) was used for unreliable data identification. Noise was estimated as an average of the first 10 frames in each utterance. The negative values in the feature vector resulting from SS were considered to be unreliable [4]:

$$|s+n| - |\hat{n}| < 0. \tag{9}$$

In the second set of experiments, additional SNR based criterion was used [11]. It treats data as unreliable if the estimated SNR is negative:

$$\hat{s}^2 < \frac{1}{2}|s+n|^2, \tag{10}$$

where $|\hat{s}| = |s+n| - |\hat{n}|$ (hat denotes estimated values).

4.3. ASR results

Figure 1 shows digit recognition accuracy as a function of SNR. Recognition is performed on the spectrally subtracted speech. The baselines are labelled 'no processing' and 'SS' and correspond to the performance of the unadapted recognizer and recognition on spectrally subtracted speech respectively. The marginalization SS+MG and state based data imputation SS+SDI techniques show similar performance. Bounded marginalization SS+BMG uses additional information (bounds) and does slightly better. Examination of the masks created via SS shows that the reliable/unreliable separation is poor and much unreliable data passes as reliable (see [11]).

Figure 2 differs from figure 1 in that an additional SNR criterion (10) is used for identification of reliable regions. The marginalization SS+SNR+MG and state based data imputation SS+SNR+SDI again show similar performance. With the SNR criterion much more data is classified as unreliable. Inspection of the errors shows that there are insertions of all models uniformly in the long runs of frames where very little data is reliable. Bounded marginalization SS+SNR+BMG performs better. Taking advantage of the known bounds, it uses the additional information and usually inserts the silence model in the unreliable data–dominated regions. Therefore, the accuracy improves considerably. Incorporation of the bounds info into data imputation is not straightforward, as discussed below. It remains to be seen if bounded data imputation can reach the level of performance of bounded marginalization.

4.4. Speech reconstruction

Figure 3 shows the potential of state based data imputation for speech reconstruction. The upper panel shows the clean speech (the string is "3162z"). The effect of telephone transmission is crudely simulated by deleting the channels 1–14 and 55–64 which are outside the 300-3000Hz range. The reconstruction proceeds in two steps. First the reduced speech is recognized and the correct state alignment is obtained. Next the missing channels are reconstructed from the means of the aligned states conditional distributions.

5. CONCLUSIONS AND FUTURE WORK

Unbounded state based data imputation performs as well as unbounded marginalization. Similar behavior was observed with Lynx helicopter and car noises. Pairing bounded marginalization with the additional SNR criterion produces a significant increase in performance. By

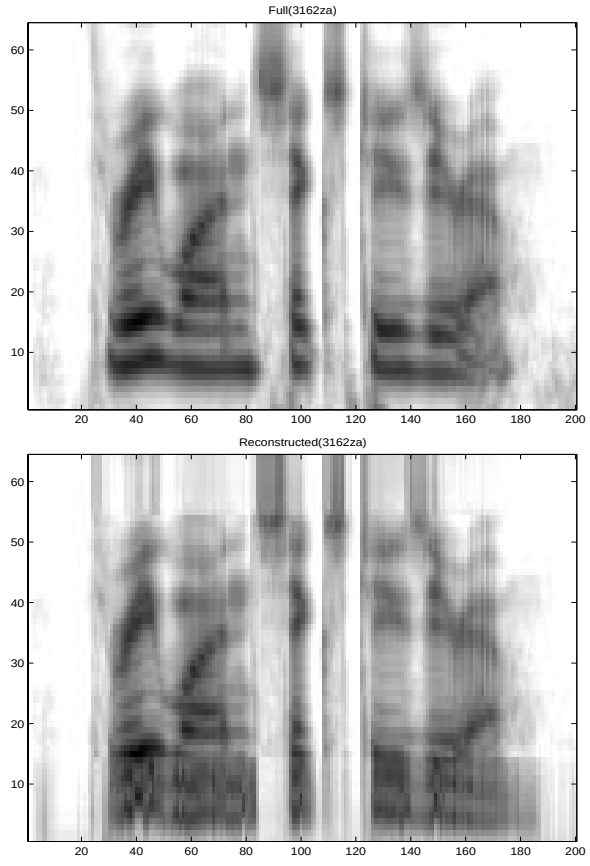

Figure 3: Full bandwidth speech (top) and reconstructed speech (bottom).

imputing values other then the mean of the conditional distribution, it might be possible to utilize the bounds constraints for data imputation, too. However, it is still unclear if this will lead to a performance level comparable to bounded marginalization.

State based data imputation opens the possibility of extending of the missing data framework to non–HMM ASR systems and/or usage of other than spectrum domain based features (e.g. cepstrum, RASTA). Since unreliable/reliable data identification happens in the spectral domain, if two sets of models are available (one of them in the spectral domain), the reliable data and spectral models can be used for an initial state alignment. Once this is obtained, the unreliable portions of the spectrum can be imputed and data transformed to the other domain for the final recognition with the second set of models. Data imputation also allows for missing data application to speech enhancement and restoration of telephone speech which, unlike ASR, requires speech reconstruction.

ACKNOWLEDGEMENTS

Thanks to Miguel A. Carreira–Perpinan for many fruitful discussions.

REFERENCES

[1] G. J. Brown and M. Cooke. Computational auditory scene analysis. *Computer speech and language*, 8:297–336, 1994.

[2] M. Cooke, P. Green, C. Anderson, and D. Abberley. Recognition of occluded speech by Hidden Markov models. Technical Report TR–94–05–01, Department of Computer Science, University of Sheffield, may 1994.

[3] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. Technical Report CS–99–05, Department of Computer Science, University of Sheffield, 1999.

[4] A. Drygajlo and M. El-Maliki. Speaker verification in noisy environment with combined spectral subtraction and missing feature theory. In *ICASSP'98*, volume 1, pages 121–124, 1998.

[5] S. Dupont. Missing data reconstruction for robust automatic speech recognition in the framework of hybrid HMM/ANN systems. In *ICSLP'98*, pages 1439–1442, 1998.

[6] P. D. Green, M. P. Cooke, and M. D. Crawford. Auditory scene analysis and Hidden Markov Model recognition of speech in noise. In *ICASSP'95*, pages 401–404, 1995.

[7] B. C. J. Moore. *An Introduction to the Psychology of Hearing*. Academic Press, 24/28 Oval Road, London NW1, 4th edition, 1982.

[8] A. C. Morris, M. P. Cooke, and P. D. Green. Some solutions to the missing feature problem in data classification, with application to noise robust ASR. In *ICASSP'98*, pages 737–740, 1998.

[9] B. Raj, R. Singh, and R. M. Stern. Inference of missing spectrographic features for robust speech recognition. In *ICSLP'98*, pages 1491–1494, 1998.

[10] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid. Speech recognition with primarily temporal cues. *Science*, 270:303–304, 1995.

[11] 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 *Eurospeech'99*, 1999.

[12] R. M. Warren, K. R. Riener, J. A. Bashford, and B. S. Brubaker. Spectral redundancy: Intelligibility of sentrences heard through narrow spectral tilts. *Perception and Psylchophysics*, 57(2):175–182, 1995.

[13] S. J. Young and P. C. Woodland. *HTK Version 1.5: User, reference and programmer manual*. Cambridge University Engineering Department, Speech Group, 1993.