MISSING DATA RECONSTRUCTION FOR ROBUST AUTOMATIC SPEECH RECOGNITION IN THE FRAMEWORK OF HYBRID HMM/ANN SYSTEMS

Stéphane Dupont[1]

Faculté Polytechnique de Mons — TCTS
31, Bld. Dolez
B-7000 Mons, Belgium
Email: dupont@tcts.fpms.ac.be

ABSTRACT

In this paper, we propose to use the missing data theory to allow the reconstruction of missing spectro-temporal parameters in the framework of hybrid HMM/ANN systems. A simple signal-to-noise ratio estimator is used to automatically detect the components that are unavailable or corrupted by noise (missing components). A limited number of multidimensional gaussian distributions are then used to reconstruct those missing components solely on the basis of the present data. The reconstructed vectors are then used as input to an artificial neural network estimating the HMM state probabilities. Continuous speech recognition experiments have been done on filtered speech. In this case, filtered components carry few or no information at all, and hence, should probably be ignored. The results presented in this paper illustrate this point of view. Complementary experiments also suggest the interest of the proposed approach in the case of noisy speech.

1. INTRODUCTION

In this paper, we propose to use the missing data theory to allow the reconstruction of damaged spectro-temporal signal portions in the framework of Automatic Speech Recognition (ASR) systems based on Artificial Neural Networks and Hidden Markov Models (hybrid HMM/ANN systems).

Contributions to the topic of robust automatic speech recognition under adverse conditions are mainly focused on two major ideas. Spectral subtraction (and derived methods like RASTA filtering [10]) allows to significantly reduce the mismatch between training and test conditions by subtracting an estimation of the noise power spectra from the spectra of the whole signal [2]. Given a noise model and the test condition, parameter compensation techniques provide a way to dynamically update the parameters of the probability density functions (pdfs) associated with the HMM states [9].

In some cases however, it could probably be better to ignore the components of the feature vectors (resulting from a filter-bank front-end) which represent spectral regions that are highly corrupted by additive noise. Moreover, components representing regions which are filtered out should also be disregarded by the subsequent classification procedure. These components are labeled as **missing** (as opposed to **present** components). Recent studies by others [7, 11] have tried to develop an automatic speech recognition architecture based on these ideas. Results show that, in some cases, up to 90% of the spectro-temporal representation can be ignored without significantly decreasing the speech recognition performance. Their work use classification based on the sampling paradigm (HMM state pdfs. are known). Gaussian Mixture Models (GMM) were used to represent the pdfs. associated with the HMM states. Clearly, this allows to compute state likelihoods on the basis of marginal distributions of the present components, hence providing a principled way to ignore feature vector components labeled as missing.

On the other hand, Artificial Neural Networks, combined with the sequence modeling capabilities of HMMs, have gained interest in the speech recognition community these past few years [4, 13]. This is probably because ANNs present an interesting alternative to GMM modeling. Indeed, ANNs allow to perform classification according to the diagnostic paradigm (based on estimations of the HMM states a posteriori probabilities). This characteristic allows to build speech recognition systems with fewer parameters than for the classical GMM systems. However, ANNs, as opposed to GMMs, do not provide any easy way to deal with missing components.

This work addresses the problem of missing components reconstruction to allow to use HMM/ANN systems. Reconstructed values are computed as the mean values of the missing components given the present components. Simple probability density functions (i.e. a limited number of multidimensional gaussian distributions) are used to model the data, thus providing a simple way to compute these conditional means. An automatic signal-to-noise ratio estimator is used to automatically detect the components that are unavailable or corrupted by noise (missing components). Reconstructed vectors are then used as input to an HMM/ANN system.

Experiments were performed for speech corrupted with additive noise severely affecting two out of fifteen critical bands. For the baseline system, the word error rate on a speaker independent telephone speech continuous numbers recognition task jumped from 11% in the case of clean speech to 60%. With the proposed approach, the degradation was more graceful: from 11% to 17%. This was slightly better than the improvement we obtained using spectral subtraction. Experiments were carried on for low-pass filtered speech with similar conclusions.

Additional experiments have been done to validate this approach in the case of wideband noise. Results were mitigated. We did not get the expected improvement. An alternative approach was also used in this case. Following [8], we used spectral subtraction [1] to enhance the corrupted speech signal prior to missing data reconstruction. Spectral subtraction lead to a significant robustness increase. Using both spectral subtraction and missing data did not yield further significant improvement.

[1] Supported by a F.R.I.A. grant (Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture), Belgium

2. MISSING DATA RECONSTRUCTION

Observation vectors x are assumed independent and identically distributed according to a probability density function made of K multidimensional gaussian distributions. The i-th distribution is characterized by the following parameters: w^i, the distribution weight, μ^i, the distribution mean and C^i, its covariance matrix. Some elements of x are labeled as missing and x can then be reorganized as follows:

$$x = (x_p x_m), \tag{1}$$

x_p for the present components and x_m for the missing components. In a similar way, we can reorganize the elements of the mean vectors and covariance matrices characterizing the pdf. of x:

$$\mu^i = (\mu_p^i \mu_m^i), C^i = \left[\begin{array}{c} C_{pp}^i C_{pm}^i \\ C_{mp}^i C_{mm}^i \end{array} \right] \tag{2}$$

We would like to reconstruct a complete vector solely on the basis of present components. Reconstruction will be done using the conditional distribution of missing components according to present components. This distribution is of the gaussian form. The reconstructed elements will be the mean of this distribution, that is to say, for the i-th gaussian:

$$x_{m|p}^i = \mu_m^i + (C_{pm}^i)^t (C_{pp}^i)^{-1} (x_p - \mu_p^i) \tag{3}$$

Considering the multi-gaussian distribution, the reconstructed value is computed as follows:

$$x_{m|p} = \frac{\sum_{i=1}^{K} w^i \phi(x_p, \mu_p^i, C_{pp}^i) x_{m|p}^i}{\sum_{i=1}^{K} w^i \phi(x_p, \mu_p^i, C_{pp}^i)} \tag{4}$$

where w^i is the weight for the i-th distribution and $\phi(x_p, \mu_p^i, C_{pp}^i)$ is the associated multidimensional gaussian distribution. This term allows to weight the contributions of the different gaussians according to the position of the present data vector in the parameter space.

An alternative approach would be to ignore the missing components and to only use the present components to compute the HMM state likelihoods on the basis of marginal distributions. For a parametric classifier (in the case of multi-gaussian HMM state modeling for instance), it is possible, although it involves a lot of computations. Indeed, covariance matrix inversions are required each time there is a change in the missing/present data configuration. This would also be possible with artificial neural network classifiers, although not really practical since this would require as many ANNs as possible data configurations.

Experiments described in [7] are related to multi-gaussian HMM state modeling. Their results are in favor of the marginal approach which yields somewhat better results than the reconstruction approach.

However, this last approach has several advantages. On the one hand, one can only use a limited number of gaussians for the reconstruction part of the system. This allows to keep a compact system (involving only a limited number of matrix inversions), without significant damage for the overall recognition system (at least during clean speech portions). On the other hand, it allows to obtain reconstructed vectors that can be used as input to any classical automatic speech recognition system. In our case, it will be an ASR system based on an ANN probability estimator.

3. SPECTRAL SUBTRACTION

Spectral subtraction was also used in this work in the case of speech corrupted with additive noise. Spectral subtraction was first used as reference. It was also applied prior to missing data reconstruction and to provide a way to identify missing components.

Generally, spectral subtraction introduces time-varying peaks and valleys in the power spectrum. These are perceived as a musical noise, which could have a significant impact on the resulting performance of the system. To reduce the effect of this noise, Berouti et al. [1] proposed a method where an overestimation of the noise is subtracted from the corrupted signal (α parameter) and where valleys are filled with a fraction of the noise power spectrum (β parameter).

Hence, spectral subtraction is implemented as follows:

$$P_o(\omega) = \left\{ \begin{array}{ll} P_s(\omega) & if P_s(\omega) > \beta P_n(\omega) \\ \beta P_n(\omega) & otherwise \end{array} \right. \tag{5}$$

$$with\ P_s(\omega) = P_i(\omega) - \alpha P_n(\omega) \tag{6}$$

$$and\ \alpha \geq 1,\ and\ 0 < \beta \ll 1 \tag{7}$$

where $P_i(\omega)$ is the corrupted input power spectrum, $P_n(\omega)$ is the noise power spectrum estimated when speech is absent from the signal, and $P_o(\omega)$ is the enhanced power spectrum. The α parameter is the overestimation factor and β is the spectral floor.

Following [8], portions where $P_s(\omega) \leq \beta P_n(\omega)$ generally corresponds to spectro-temporal regions where noise dominates and can then be identified as missing for the subsequent reconstruction module.

4. EXPERIMENTS ON THE NUMBERS'93 CORPUS

Experiments have been done on a telephone speech connected numbers recognition tasks. We have been using the NUMBERS'93 [6] database. The training set was composed of 1534 utterances (of which 1400 were used to adjust the weights of the ANNs and 134 for cross-validation purposes) and 384 utterances were used for testing. The corpus is composed of 36 vocabulary words and the *CMU 0.4* lexicon was used to obtain phonetic transcriptions for these words.

A HMM/ANN hybrid baseline system was first developed. Acoustic feature vectors were log-RASTA [10] filtered critical band energies (complemented by their first derivatives). Log-RASTA processing yields increased robustness towards channel variabilities. Moreover, 9 feature frames were used as input to the system, providing the ANN with contextual information. The ANN was a feedforward multilayer perceptron. The word error rate of this system was **13.6%**. It should be noted here that we will have to operate on spectral parameters. Indeed, we want to reconstruct spectro-temporal portions that are filtered out or corrupted by additive noise. A spectral representation of the speech is used as it allows to isolate (and label as missing) the corrupted frequency regions.

Then, we have developed a system based on the described reconstruction approach. As explained in the subsequent sections, several configurations have been investigated regarding the reconstruction model. The module in charge of the missing portions detection was either based on a signal-to-noise ratio estimator using energy histograms within the different frequency channels [3] or else on the spectral subtraction approach (in the case of speech corrupted with wideband noise).

4.1. Reconstruction models

4.1.1. Diagonal Covariance Matrices

We have been using multi-gaussian distributions with diagonal covariance matrices. In this case, reconstruction is nearly immediate because the second term of Equation 3 is null. The reconstruction simply operates by replacement of the missing data by their unconditional distribution mean. The drawback of this approach is that it does not allow to use the correlation between the feature vector elements, which is important between adjacent frequency channels. And indeed, using a mono-gaussian distribution did not yield any improvement. On the contrary, some degradation was observed, probably because the distortion introduced by the reconstruction was higher than the distortion resulting from filtering (see Figure 1).

Multi-gaussian distributions were used with some success. They allow to model more complex distribution forms and can capture correlation between feature elements. As can be seen in Figure 1, this configuration yielded significant performance improvement in extreme filtering cases and for distributions made of a large number of gaussians. This approach has also been tested in [7]. Results showed poor performances.

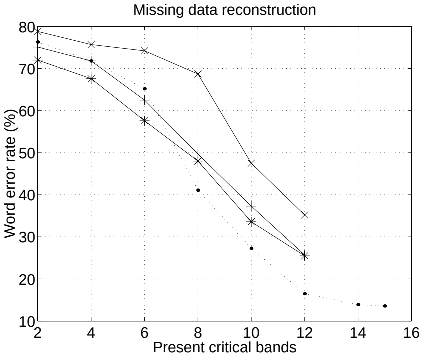

Figure 1. Word error rates for missing data reconstruction using diagonal covariance matrices. The dotted line is for the baseline system. The '×','+' and '∗' lines respectively correspond to distributions modeled with 1, 16 and 64 gaussians.

4.1.2. Full Covariance Matrices

Full covariance matrices were used with more success. This is justified by the strong correlation between the different spectral components of the acoustic vectors. However, this kind of modeling requires a lot of matrix inversions, which could become prohibitive. Results presented on Figure 2 already show a highly significant word error rate decrease with a distribution modeled by a single multidimensional gaussian. Using more gaussians (32) only yielded a marginal improvement, in the case of low-pass filtering at least.

4.1.3. Using Context Frames

Diagonal as well as full covariance models were also used to model several adjacent feature frames. Correlation across time could indeed help the reconstruction of missing elements.

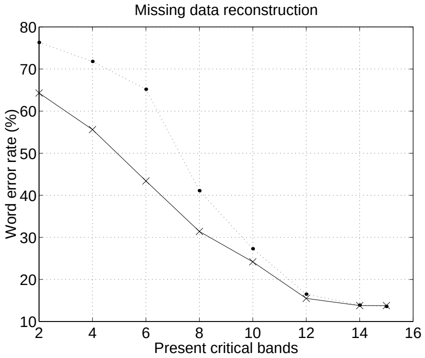

Figure 2. Word error rates for missing data reconstruction using full covariance matrices. The dotted line is for the baseline system and the continuous line is for reconstruction system using 1 gaussian.

Several configurations were tested. These involved from 2 up to 232 (4*58 HMM states) gaussians, modeling the distribution of 1, 3 or 5 adjacent spectral frames, with or without their first and second derivatives. In the case of low-pass filtered speech, these experiments indicated a slight preference for a system using 232 gaussians and a single feature frame. Using a wider acoustic context generally did not yield improvements, while possibly leading to resource overload. This, however, does not preclude the use (and interest) of context distribution modeling, which could reveal its potential for other kinds of perturbations like interruptions and wideband noises.

The complexity of such an approach could be reduced by using common covariance matrices. Additionally, the size of the feature vectors could be reduced by selecting time-frequency points distributed across the spectro-temporal plane.

4.2. Noise-Corrupted Speech

Then, we investigated the case of a signal corrupted with additive sine wave. Results are presented in Figure 3. It clearly shows that the system based on missing data reconstruction outperforms the other systems, including the systems based on spectral subtraction and on multiband recognition [5]. The excellent performance can be explained by the fact that only two (because of the slight overlap between critical band filters) out of fifteen critical bands are corrupted. Hence, missing components can reliably be reconstructed on the basis of the other parameters and their first derivatives.

We finally investigated the case of a wideband noise perturbator. Speech was corrupted with gaussian white noise at different signal-to-noise ratios. In this case, regions with high speech-energy, and which dominate the noise, are identified as present, while low-energy regions are labeled as missing. This gives rise to missing data detection patterns which are radically different from those obtained in the case of coloured noise. As can be expected, we obtain spectral patterns isolating formants and high-energy vocalic regions from the remainder of the time-frequency representation. Missing data reconstruction based on this pattern did not yield the expected improvement. We even observed some additional degradation as we increased the minimum SNR threshold for considering a frequency channel as present.

Spectral subtraction was then used as pre-processing and missing data identification scheme. Results are summarized in Table 1. As can be seen, we did not get any significant improvement by com-

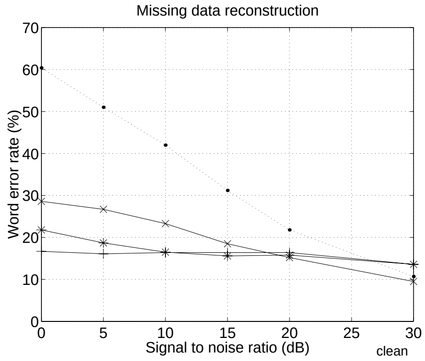

Figure 3. Word level error rates for different noise levels (400 Hz sine wave, global SNR from 0 dB to 30 dB). The dotted line is for the baseline system. The '$+$', '$\times$' and '$*$' line are respectively for the missing data reconstruction approach (based on a mono-gaussian distribution and a single feature frame), for the multiband approach and for the spectral subtraction approach.

bining spectral subtraction and missing data reconstruction.

It should be noted that all these experiments were done using a mono-gaussian full covariance matrix reconstruction model. Although this model performed well in the case described in the previous section, it may not be accurate enough in the current case. Work in progress particularly investigates the use of more complex distributional modeling as well as time-domain correlation across feature frames.

Error Rate (%)	0 dB	10 dB
Baseline	76.1	32.7
Spectral Subtraction	50.5	30.8
Spectral Subtraction + Missing Data	49.5	30.3

Table 1. Word error rates on continuous numbers recognition (NUMBERS'93 database) with white noise addition (0 dB and 10 dB SNR). Reconstruction was based on a full covariance matrix mono-gaussian distribution and a single feature frame.

5. DISCUSSION AND CONCLUSIONS

This paper presents the use of a missing data reconstruction technique in the framework of automatic speech recognition systems using Artificial Neural Networks together with Hidden Markov Models.

Results presented here show that a fairly simple reconstruction system, controlled by an elementary signal-to-noise ratio estimator, provides a significant robustness improvement in the case of low-pass filtering as well as coloured additive perturbations. White noise perturbations were then investigated without success. Missing data reconstruction alone, as well as combined with spectral subtraction did not yield the expected improvement. Further work includes the use of several context frames and more accurate reconstruction models.

Applying missing data, as well as spectral subtraction, requires to work in the spectral domain. Our experiments were done with critical band energies. Linear prediction cepstral coefficients could also be computed from the reconstructed critical band energies. Generally, cepstral coefficients perform better that straight spectral coefficients. Recent work in the field of multiband speech recognition [5] has shown that the use of linear prediction cepstral coefficients computed on narrow frequency bands also yields an inherent robustness compared to the use of straight critical bands and to the use of LP cepstral coefficients computed on the whole frequency range. Experiments by us [5] and by others [14, 12] confirm this observation for different frequency band configurations. Parameters of this kind could also be used.

REFERENCES

[1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. of ICASSP'79*, pp. 208–211, Apr. 1979.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE ASSP*, vol. 2, no. 27, 1979.

[3] H. Bourlard, S. Dupont, H. Hermansky, and N. Morgan, "Towards sub-band-based speech recognition," in *Proc. of European Signal Processing Conference*, (Trieste, Italy), pp. 1579–1582, Sept. 1996.

[4] H. Bourlard and N. Morgan, *Connectionist Speech Recognition – A Hybrid Approach*. Kluwer Academic Publishers, ISBN 0-7923-9396-1, 1994.

[5] H. Bourlard and S. Dupont, "Sub-band-based speech recognition," in *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, (Munich), pp. 1251–1254, Apr. 1997.

[6] R. Cole, M. Fanty, and T. Lander, "Telephone speech corpus at cslu," in *Proc. of Intl. Spoken Language Processing*, (Yokohama, Japan), September 1994.

[7] M. Cooke, A. Morris, and P. Green, "Missing data techniques for robust speech recognition," in *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, (Munich), Apr. 1997.

[8] M. El-Maliki, P. Renevey, and A. Drygajlo, "Rehaussement par soustraction spectrale et compensation des paramtres manquants pour la reconnaissance robuste du locuteur et de la parole," in *Proc. XXIImes Journes d'Etude sur la Parole*, (Martigny, Switzerland), pp. 409–412, 1998.

[9] M. Gales, "Nice model-based compensation schemes for robust speech recognition," in *Proc. of ESCA/NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 55–64, Apr. 1997.

[10] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[11] R. P. Lippmann and B. A. Carlson, "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise," in *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, (Munich), pp. KN37–KN40, Apr. 1997.

[12] S. Okawa, E. Bocchieri, and A. Potamianos, "Multi-band speech recognition in noisy environments," in *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, (Seattle, WA), pp. 641–644, 1998.

[13] T. Robinson, M. Hochberg, and S. Renals, "The use of recurrent networks in continuous speech recognition," in *Automatic Speech and Speaker Recognition - Advanced Topics*, pp. 233–258, Kluwer Academic Publishers, 1996.

[14] S. Tibrewala and H. Hermansky, "Sub-band based recognition of noisy speech," in *Proc. of ICASSP'97*, (Munich), pp. 1255–1258, 1997.