Experiments on Hiwire database using Denoising and Adaptation with a hybrid HMM-ANN Model

Roberto Gemello [1], *Franco Mana* [1], *Stefano Scanzio* [2]

[1] Loquendo, Torino, Italy
[2] Politecnico di Torino, Italy

roberto.gemello@loquendo.com, franco.mana@loquendo.com, stefano.scanzio@polito.it

Abstract

This paper presents the results of a large number of experiments performed on the Hiwire cockpit database with a hybrid HMM-ANN speech recognition model[1]. The Hiwire database is a noisy and non-native English speech corpus for cockpit communication. The noisy component of the database has been used to test two noise reduction methods recently introduced, while the adaptation component is exploited to perform supervised and unsupervised adaptation of the HMM-ANN model with an innovative technology, both in multi-speaker and speaker dependent way. Baseline results are presented, and the improvements obtained with noise reduction and adaptations are reported, showing an error reduction of about 60%.

1. Introduction

The Hiwire database (HDB) is a noisy and non-native English speech corpus for cockpit communication [1], recently made available to the speech recognition community by the Hiwire Consortium. The database includes short vocal sentences in English, corresponding to aeronautic commands. It has been recorded by 81 non-native speakers from 4 countries: France, Greece, Italy and Spain. Speech data have been recorded in quiet rooms with a close-talking microphone (Plantronics USB-45) and mixed with real noise recorded into a Boeing 737 cockpit with a boundary microphone (AKG Q300).
The noise has been mixed at different levels to obtain three SNRs: 10, 5 and -5dB, conventionally named LN (low noise), MN (medium noise) and HN (high noise).
This paper employs the HDB in three ways:

1. to establish baseline results using a state-of-art hybrid HMM-ANN recognition model (Loquendo ASR), employing both the native 16 kHz sampling rate, and down-sampling to telephonic band (8 kHz);
2. to produce improved results using a new noise reduction method recently introduced in [4], comparing it with a more classical method;
3. to test a new adaptation method for HMM-ANN, already introduced in [6], both in single speaker and in a multi-speaker/multi-condition adaptation mode.

2. Loquendo Hybrid HMM-ANN model

The Loquendo-ASR system uses acoustic models based on a hybrid combination of Hidden Markov Models (HMM) and Multi Layer Perceptron (MLP) [2], where each phonetic unit is described in terms of a single or double state left-to-right automaton with self-loops, the HMM transition probabilities are uniform and fixed, and the emission probabilities are computed by an MLP [3]. The MLP is characterized by a wide input window modelling a time context of 210 ms (21 frames), fed by a set of Rasta-PLP parameters, including log Energy and 12 cepstrals, together with their first and second derivatives, for a total of 819 units. The first hidden layer is divided into three blocks, one for the central frames, and two for the left and right context (200 units each). Each block is, in its turn, divided into six sub-blocks devoted to keep into account the six types of different input parameters. This structured layer was empirically found to be better than a fully connected one. This first layer is followed by two fully connected hidden layers of 300 units. Using three hidden layers, rather than a larger single hidden layer, has the advantage of reducing the total number of connections, and improves also the performance.

The output layer estimates the posterior probabilities of the acoustic units, which are based on a set of vocabulary and gender independent units including stationary context-independent phones and diphone-transition coarticulation models [3]. The complete network structure, for US English, has 5 levels, with dimensions 819-600-300-300-949 from input to output.

3. Denoising Methods

As the HDB is a noisy corpus, it is necessary to employ denoising methods to improve the baseline system performance. Two methods, developed within the Hiwire project, have been employed in this work. They are fully described in [4], and briefly summarized in the following.

3.1. SNR Dependent Wiener Spectral Subtracton

Let $|Y_k(m)|^2$, $|X_k(m)|^2$ and $|D_k(m)|^2$ be the k-th frequency sample of the spectrum energy of noisy speech, clean speech, and noise respectively.
SNR dependent Wiener Spectral Subtraction [4] is defined by:

$$|X_k(m)|^2 = \begin{cases} \dfrac{\left[|Y_k(m)|^2 - \alpha(m)|\hat{D}_k(m)|^2\right]^2}{|Y_k(m)|^2} & \text{if } |Y_k(m)|^2 - \alpha(m)|\hat{D}_k(m)|^2 > \beta(m)|Y_k(m)|^2 \\ \beta(m)|Y_k(m)|^2 & \text{otherwise} \end{cases} \quad (1)$$

where $\alpha(m)$ is a noise overestimation factor, and $\beta(m)$ is a spectral floor used to avoid negative spectrum values. These two parameters vary in time as function of the Signal-to-Noise Ratio SNR(m), computed as follows:

[1] This work was supported by the EC FP-6 IST Project HIWIRE – Human Input that Works in Real Environments

$$SNR(m) = 10\log_{10}\left(\frac{\sum_k |X_k(m)|^2}{\sum_k |\hat{D}_k(m)|^2}\right) \quad (2)$$

where $|\hat{D}_k(m)|^2$ is an estimation of the k-th noise spectral row at time m; $\alpha(m)$ and $\beta(m)$ are defined as possibility functions of SNR(m) as shown in Fig. 1.

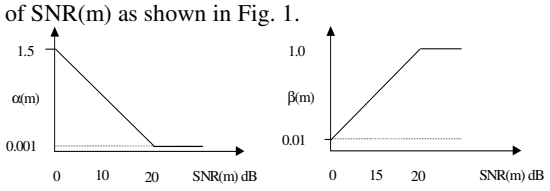

Fig. 1. *Definition of α(m) and β(m) as functions of SNR(m)*

3.2. SNR Dependent Ephraim-Malah noise suppression

Ephraim–Malah MMSE log estimator is a short-time spectral amplitude estimator that minimizes the mean-square error of the estimated logarithms of the spectra. It is defined as follows:

$$G_k = \frac{\xi_k(m)}{1+\xi_k(m)}\exp\left(\frac{1}{2}\int_{\nu_k(m)}^{\infty}\frac{e^{-t}}{t}dt\right) \quad (3)$$

where:

$$\xi_k(m) = \frac{|X_k(m)|^2}{|D_k(m)|^2} \text{ is the } a\ priori \text{ SNR,} \quad (4)$$

$$\gamma_k(m) = \frac{|Y_k(m)|^2}{|D_k(m)|^2} \text{ is the } a\ posteriori \text{ SNR,} \quad (5)$$

and $\nu_k(m) = \frac{\xi_k(m)}{1+\xi_k(m)}\gamma_k(m)$

The computation of the *a priori* SNR requires the knowledge of the clean speech spectrum, which is not available. Its estimation can be obtained with *a decision-directed approach* as follows:

$$\hat{\xi}_k(m) = \eta(m)\frac{|\hat{X}_k(m-1)|^2}{|\hat{D}_k(m-1)|^2} + [1-\eta(m)]\max[0, \gamma_k(m)-1] \quad (6)$$
$$\eta(m) \in [0,1)$$

In [4] it was found to be convenient, for speech recognition applications, to make the estimation of the *a priori* and the *a posteriori* SNR dependent on the noise overestimation factor $\alpha(m)$ and on the spectral floor $\beta(m)$ as follows:

$$\tilde{\xi}_k(m) = \max\left(\eta(m)\frac{|\hat{X}_k(m-1)|^2}{\alpha(m)|\hat{D}_k(m-1)|^2} + (1-\eta(m))[\tilde{\gamma}_k(m)-1], \beta(m)\right) \quad (7)$$
$$\eta(m) \in [0,1)$$

$$\tilde{\gamma}_k(m) = \max\left(\frac{|Y_k(m)|^2}{\alpha(m)|\hat{D}_k(m)|^2} - 1, \beta(m)\right) + 1 \quad (8)$$

where the noise overestimation factor $\alpha(m)$ and the spectral floor $\beta(m)$ are function of SNR(m) as shown in Fig. 1. Our approach modifies the estimation of γ_k and ξ_k while maintaining the global shape of the gain function $G_k(\gamma_k, \xi_k)$. The modified gain function can be expressed as follows:

$$\tilde{G}_k(\gamma_k(m), \xi_k(m)) = G_k(\tilde{\gamma}_k(m), \tilde{\xi}_k(m))$$

with $\tilde{\xi}_k(m), \tilde{\gamma}_k(m)$ computed according to (7) and (8).

4. Adaptation methods

The presence of non-native speakers in the HDB is a serious problem that affects system performance. In the Hiwire project, two approaches have been used to deal with this problem:

- A linguistic approach, using additional transcriptions and language dependent phoneme mapping;
- An acoustic model adaptation approach: the one considered in this paper.

In this paper, model adaptation has to be applied to a Hybrid HMM-ANN model. While the literature on adaptation is rich of techniques for refining ASR systems by adapting the acoustic features and the parameters of stochastic models (HMM Gaussian mixtures), far less proposals have been made for the adaptation of the features and model parameters of systems that use the hybrid HMM-ANN approach.

4.1. Input Feature Transformations

The simplest and more popular approach to speaker adaptation with ANNs is Linear Input Transformation (LIN) [5]. The input space is rotated and shifted by a linear transformation to make the target conditions more consistent with the training conditions. The LIN weights are initialized with an identity matrix, and trained by minimizing the error at the output of the ANN keeping fixed the weights of the original ANN.

4.2. Hidden feature transformations

This new technique has been introduced in the Hiwire project and described in [6]. Assuming that the activation values of a hidden layer represent an internal structure of the input pattern in a space more suitable for classification, a linear transformation is applied to the activations of an internal layer. Such a transformation is performed by a Linear Hidden Network (LHN). The values of an identity matrix are used to initialize the weights of the LHN. The weights are trained using a standard back-propagation algorithm keeping frozen the weights of the original network. Since the LHN performs a linear transformation, once the adaptation process is completed, the LHN can be removed combining LHN weights with the ones of the next layer.

A method for avoiding the forgetting of acoustic-phonetic classes that are not represented in the adaptation set (Conservative Training [6]) is also employed in the experiments.

5. Experiments

Three sets of experiments have been carried on the HDB by using Loquendo ASR.

The first one was devoted to obtaining baseline results both with telephone (8 kHz) and microphone (16 kHz) models. The second set illustrates the improvements obtained with the denoising methods. The third one reports the improvement reached with acoustic model adaptation, alone and in combination with denoising.

5.1. Experimental Conditions

5.1.1. Training

Two HMM-ANN models have been trained:

- ***Telephone 8 kHz***: trained with a large telephone corpus (LDC Macrophone + SpeechDat Mobile)
- ***Microphone 16 kHz***: trained with a collection of microphone corpora (timit, wsj0-1, vehic1us-ch0)

5.1.2. Adaptation

The first 50 utterances of each speaker have been used for adaptation, pooling all the noise conditions: Clean, LN (10dB SNR), MN (5dB SNR), HN (-5dB SNR).
The seed model for adaptation was *Microphone 16 kHz*.
Two kind of adaptation were performed:

- **Multi-Condition**: the adaptation data of all the speakers and all noise conditions is pooled. The models are adapted to channel, noise conditions, and non-native common aspects.
- **Speaker-Dependent**: Adaptation and tests are performed for each speaker separately, and all results are finally averaged. The models are adapted mainly to speaker's voice, but also to channel and noise conditions.

LHN adaptation method is employed, with two variants:

- Conservative Training on (LHN cons), that performs adaptation preserving the acoustic-phonetic classes not represented in the adaptation set [6].
- Conservative Training off (LHN spec), that specializes completely the models to the adaptation set.

Experiments with two types of adaptation are performed:

- **Supervised**: the transcriptions of the sentences available in HDB are employed to perform forced segmentation of the adaptation utterances, providing the labels needed by the adaptation process, which is intrinsically supervised.
- **Unsupervised**: the transcriptions of the sentences are not employed, to simulate an "on-the-field" adaptation, and are approximated by the ASR outputs. Only the adaptation utterances recognized with a certain degree of confidence are used in the adaptation process, to avoid divergence due to incorrectly labeled data.

5.1.3. Test

The last 50 utterances of each speaker are used for testing.
The standard aeronautic commands grammar made available by Thales-Avionique (and obtainable from the Hiwire consortium) has been employed.
Examples of typical sentences are:

- Request Descent to five five two
- FD Engage
- ETA zero victor five nine whiskey At six nine

5.2. Results

The experimental results have been collected into 3 Tables. All results are given as Word Accuracy (%WA). Error Reduction percentage (%E.R.) is reported with respect to the baseline results.

5.2.1. Baseline and Denoising

Table 1 shows baseline and denoising results. The average baseline results are 42.4% for *Telephone 8 kHz* and 43.0% for *Microphone 16 kHz*. The *16 kHz* models are more accurate on clean speech (90.5% vs. 88.4%), but the *8 kHz* models are more robust to noisy conditions, due to the larger presence of real noise in the training corpus. Ephraim-Malah noise reduction always outperforms Wiener spectral subtraction (32.8% vs. 25.7% and 25.7% vs. 21.8% E.R.). The *Microphone 16 kHz* model is used as baseline in Tab. 2-3 and the %E.R. is always computed referring to it.

5.2.2. Adaptation

5.2.2.1 Multi-Condition

The results of Multi-Condition adaptation are reported in Table 2, and are divided into three sections: the first and second sections show the results for supervised adaptation, without and with denoising respectively. The third section reports the results for unsupervised adaptation with denoising. The results show that:

- supervised multi-condition adaptation gives good performance improvement. It operates well even without denoising, since it incorporates information of channel, noise and non-native accents in the models. 98.2% WA is the best reported performance on the clean part of the Hiwire database.
- The average best results are obtained with supervised adaptation in conjunction with denoising (60.7% E.R.) The non-conservative method (LHN spec) gives better results because there is complete coherence between the adaptation and test sets. The reduction of generalization capability of the adapted models is not revealed by this test.
- As expected, unsupervised adaptation is inferior to supervised adaptation (51.7% vs. 60.7% E.R.), but it proves to be an effective technique for adaptation in real life applications, when transcriptions of vocal material are not available. In that case, the conservative method helps to contain performances degradation (51.7% vs. 47.4% E.R.).

5.2.2.2 Speaker Adaptation

The results of speaker adaptation are shown in Table 3. Speaker adaptation is very effective for improving performances on HDB. The error reduction achieved by Supervised Adaptation plus Ephraim-Malah noise reduction is very large (56.3%) increasing WA on clean data from 90.2 to 95.4, and showing even larger improvement in noisy conditions (in the HN condition the WA is more than doubled, from 16.6 to 33.7). The limitation for this approach is that it suffers the scarcity of adaptation data (~150s of voice per speaker). The Unsupervised Adaptation obtains a 49.6% error reduction. According to our expectations, the error reduction is inferior to supervised adaptation, due to the errors introduced by the ASR transcriptions, but still it is relevant.

6. Conclusions

The HDB has been processed to produce baseline results, with a large HMM-ANN model, both with a *16 kHz microphone* model and a *8 kHz telephone* model. Two denoising methods recently introduced in the Hiwire project have been tested, producing improved results. The standard HDB adaptation component has been used to perform several acoustic model adaptation experiments, including supervised and unsupervised, single speaker and multi-speaker adaptations. The results obtained compares favorably with the previously published results on the same corpus with other methods.

7. References

[1] J.C. Segura, T. Ehrette, A. Potamianos, D. Fohr, I. Illina, P-A. Breton, V. Clot, R. Gemello, M. Matassoni and P.

Maragos, "The HIWIRE database, a noisy and non-native English speech corpus for cockpit communication", online at http://www.hiwire.org/, 2007.
[2] H. Bourlard, N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach,* Kluwer Academic Publishers, 1993.
[3] D. Albesano, R. Gemello, and F. Mana, "Hybrid HMM-NN Modelling of Stationary-Transitional Units for Continuous Speech Recognition", Int. Conf. On Neural Information Processing, pp. 1112–1115, 1997.
[4] R. Gemello, F. Mana and R. De Mori, "Automatic Speech Recognition with a Modified Ephraim-Malah Rule", IEEE Signal Processing Letters, vol. 13, no. 1, pp. 56-59, January 2006
[5] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, and S. Renals, T. Robinson, "Speaker-adaptation for Hybrid HMM-ANN Continuous Speech Recognition System," Proc. EUROSPEECH 1995, pp. 2171–2174, 1995.
[6] D. Albesano, R. Gemello, P. Laface, F. Mana, S. Scanzio, "Adaptation of Artificial Neural Networks Avoiding Catastrophic Forgetting", Proc. Of Int. Joint Conference on Neural Networks 2006, Vancouver, Canada.

Table 1: *Results for the HDB with a telephone 8 kHz and a microphone 16 hKz model. Results are reported without denoising (No Den) and with two denoising methods (SNR dep. Wiener - WIE and Ephraim-Malah - EM), and as a function of the noise condition: Clean, LN (10dB SNR), MN (5dB SNR), HN (-5dB SNR), and AVG (Average of all conditions).*

Models	Denoising method	Noise Condition				AVG	E.R. %
		Clean	LN	MN	HN		
Telephone 8 kHz (Macrophone)	No Den	88.4	51.1	27.3	2.8	42.4	-
	WIE	88.3	70.0	54.1	16.3	57.2	25.7
	EM	88.3	74.7	62.0	20.1	61.3	32.8
Microphone 16kHz (timit-wsj0-1-vehic1us)	No Den	90.5	49.1	27.5	5.0	43.0	-
	WIE	90.4	68.5	51.1	14.5	56.2	23.2
	EM	90.2	71.9	55.0	16.6	58.4	27.0

Table 2: *Results for the HDB using multi-condition adaptation. The adaptation sets of all speakers in all noise conditions are pooled together. Results are shown separately for adaptation alone, and in conjunction with SNR dep. Ephraim-Malah noise reduction. Two different types of adaptation are performed: Supervised (Supv), where the transcription of the sentences is available, and Unsupervised (Unsupv), where the transcription of the sentences is not available, and is approximated by an ASR generated transcription. Furthermore, two variants are tested: with Conservative Training (LHN cons) and without it (LHC spec).*

Multi-Condition Adaptation		Denoising method	Noise Condition				AVG	E.R. %
Method	Type		Clean	LN	MN	HN		
No	-	No	90.5	49.1	27.5	5.0	43.0	-
LHN cons	Supv		97.5	81.1	59.2	13.4	62.8	34.7
LHN spec	Supv		98.2	90.9	79.6	34.8	75.9	57.7
No	-	EM	90.2	71.9	55.0	16.6	58.4	27.0
LHN cons	Supv		97.1	90.6	79.3	31.1	74.5	55.3
LHN spec	Supv		98.0	93.2	83.7	35.5	77.6	60.7
LHN cons	Unsupv	EM	94.3	87.2	76.8	31.5	72.5	51.7
LHN spec	Unsupv		93.7	85.5	73.7	27.1	70.0	47.4

Table 3: *Results for the HDB using speaker adaptation. Adaptation and test is performed for each speaker separately, and then all results are averaged. Two different types of adaptation are performed: Supervised, where the transcription of the sentences is available, and Unsupervised, where the transcription of the sentences is not available, and is approximated by an ASR generated transcription. SNR dep. Ephraim-Malah noise reduction is always applied. Results are shown separately for each noise condition: Clean, LN (10dB SNR), MN (5dB SNR), HN (-5dB SNR), and AVG (Average of all conditions).*

Speaker Adaptation		Noise Condition				AVG	E.R. %
Method	Type	Clean	LN	MN	HN		
No	-	90.2	71.9	55.0	16.6	58.4	27.0
LHN cons	Supv	95.4	90.1	81.2	33.7	75.1	56.3
LHN cons	Unsupv	93.7	85.8	74.8	30.7	71.3	49.6