IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

Error Correcting Posterior Combination for Robust Multi-Band Speech Recognition

Astrid Hagen [§] Hervé Bourlard [§]

IDIAP–RR 01-10

March 2001

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

§ IDIAP—Dalle Molle Institute of Perceptual Artificial Intelligence, P.O. Box 592, CH-1920 Martigny, Switzerland, {hagen,bourlard}@idiap.ch.

IDIAP Research Report 01-10

Error Correcting Posterior Combination for Robust Multi-Band Speech Recognition

Astrid Hagen Hervé Bourlard

March 2001

Abstract. In human perception, the availability of context enhances recognition and renders it more robust to noise. Even if not all phonemes in a word (or words in a sentence etc.) are correctly perceived, humans can fill in missing parts with the help of cues from the surrounding speech parts. This was proven in studies on human speech perception where recognition of words in sentences under noise was shown to outperform recognition of words in isolation or, even more drastically, of nonsense syllables under noise.
A new model for quantifying the influence of contextual information on human recognition performance was recently proposed. Although the authors state that it is not a model for the recognition process itself, we will see how the ideas behind this model can be used in automatic speech recognition to extend our formerly introduced multi-band recognition systems to incorporate frequency contextual information. We will compare the new set-up to our former models such as the full combination subband approach and its approximation.

Acknowledgements: This work was supported by the Swiss Federal Office for Education and Science (OFES) in the framework of both the EC/OFES SPHEAR (SPeech, HEAring and Recognition) project and the EC/OFES RESPITE project (REcognition of Speech by Partial Information TEchniques).

1 Introduction

Several approaches to render automatic speech recognizers more robust against corruptions by noise exist. One of the most promising approaches is the use of contextual information which can account for local errors introduced by the noise. In automatic speech recognition (ASR) systems, contextual information *over time* is, thus, usually included (up to a certain degree) through the use of larger time-scale information which is incorporated e.g. by a long input window in the case of MLP recognizers and/or by the use of delta and possibly delta-delta features. Another (additional) approach to render speech recognizers more noise robust is multi-band processing. In multi-band based recognizers, the frequency domain is split into several frequency subbands which are processed separately for feature extraction and probability estimation before the estimates from all subbands are recombined for decoding. For this reason, multi-band systems can miss important correlation over frequency. An approach which tries to overcome this limitation explicitly models all possible combinations of subbands [3, 5] and has shown improved noise robustness as compared to original multi-band processing.

We will see in the following, how this "full combination" (FC) approach to subband processing can be further improved by introducing an additional processing step inspired from a "model for context effects in human speech recognition" described in [1]. In Section 2, we will briefly discuss this model as proposed by [1][1] and show the similarity of its formalism to the FC approach and its approximation. The approximation to FC (AFC) is employed when a high number of subbands is to be used and the training of one recognizer for every possible combination of subbands, as it is done in FC, becomes unfeasible. We will then show how the two multi-band approaches (FC and AFC) can be enriched through application of the human model to automatic processing. In Section 5, the experiments which were carried out in the framework of HMM/ANN (Hidden Markov Model/Multilayer Perceptron) hybrid systems on clean and noise-corrupted data will be presented.

2 The Models for Context Effects

In this section, we briefly introduce the two-stage model of human perception as proposed by Bronkhorst, Bosman and Smoorenburg [1] and show how it could be interpreted in the framework of a multi-band recognizer for ASR.

2.1 A two-stage model of human perception

Bronkhorst et al.'s [1] model is based on a description of human perception as a two-stage process: A listener first tries to identify the stimulus by using sensory information only. Then, in the second part, he/she increases the accuracy of the perceived stimulus by filling in the missing parts of (an incompletely perceived) stimulus by the use of contextual information.

2.1.1 Measure of sensory information

The probability of occurrence of a (possibly incomplete) stimulus is calculated from the probabilities of correct recognition of the n elements in the stimulus, which are denoted by p_i, $i=1..n$. For recognition of an element and, thus, the calculation of p_i, only the ('sensory') information contained in the element itself is used. Thus, it is implicitly assumed that the recognition probabilities of the elements are independent at this stage, and that the probability of identifying the whole stimulus can be described as the product of the element recognition probabilities. As each element can either be recognized correctly p_i or incorrectly $1-p_i$, the overall recognition probability for the whole stimulus is described by all possible combinations Q_i $(i = 0..n)$ of $(n-i)$ correctly and i erroneously identified element

[1]This model describes the use of context information over time. We will see how its strategy can also easily be used over frequency.

probabilities, i.e.:

$$\begin{aligned}
Q_0 &= p_1 p_2 ... p_n, \\
Q_1 &= (1-p_1)p_2...p_n + ... + p_1...p_{n-1}(1-p_n) \\
&\vdots \\
Q_n &= (1-p_1)(1-p_2)...(1-p_n)
\end{aligned}$$

where each Q_i consists of $\binom{n}{i}$ terms which represent all possible permutations of i missing elements in the set of n elements. (For all Q_i together, there are $N = 2^n$ terms, i.e. combinations of present and missing elements.) Q_i is referred to as a 'percept'.

2.1.2 Measure of contextual information

It is assumed that a listener has the chance c_i of correctly guessing *exactly* one of i missing elements in a stimulus, excluding the possibility that more than one element is corrected at a time. The elements which are correctly guessed are then assumed to correspond to correctly perceived elements from the first stage, i.e. a corrected element is regarded as if it had not been wrongly perceived in the first place. Thus, a listener correctly perceiving $(n-i)$ elements has, each time, a chance $c_i c_{i-1}...c_1$ of correctly guessing the missing elements: c_i for correcting the i^{th} element, c_{i-1} for correcting the $(i-1)^{th}$ element, and so forth. The parameter c_i therefore describes the influence of context in the recognition process.

There are different possibilities for estimating the context parameters c_i. In word recognition, e.g., for each number i of missed phonemes the alternative words are counted in the lexicon. These numbers together with the total number of words in the dictionary can then be used to calculate each c_i.

An estimate of the average correct recognition probability $p(w)$ of the whole stimulus can then be established by multiplying the probability of occurrence of a certain percept Q_i by the chance of guessing the stimulus, and adding over all possible percepts:

$$p(w) = Q_0 + c_1 Q_1 + c_1 c_2 Q_2 + ... + c_1 ... c_n Q_n \tag{1}$$

If all c_i were assumed equal to c (i.e. $c^i = \prod_{j=1}^{i} c_j$ with $c_j = c$), (1) can be written in a closed form:

$$p(w) = \sum_{i=0}^{n} c^i Q_i \tag{2}$$

Testing the model with different kinds of (non-equal) estimates of c_i, Bronkhorst et al. showed that recognition scores of consonant-vowel-consonant (CVC) words in auditory or orthographic presentation could be well predicted with this model.

2.2 Application to Subband-based ASR

In the following, we discuss how to use and extend the above formalism to subband-based ASR and, thus, to introduce frequency contextual information. The probability of correct recognition of an element now corresponds to the probability of correct recognition of a reliable time-frequency block, and the probability of erroneous recognition of an element corresponds to the probability of incorrect recognition of an unreliable block (unreliable due to corruption by noise).

2.2.1 Measure of local information

In subband-based ASR, elements are now interpreted as regions in the frequency domain belonging to one time frame, i.e. as time-frequency blocks. An element probability $p_i = P(q_k|x_i)$ of correctly recognizing an element, thus, corresponds to the (posterior) probability of correct recognition of that frequency block (for phoneme q_k, ($k = 1..K$)) by the subband recognizer which was trained on this frequency region i ($x_i =$ frequency subband i of x, $i = 1..n$, for one time frame). The probability of incorrect recognition is therefore $1-p_i = 1-P(q_k|x_i)$.

A certain recognition event ('percept') can be described as a combination of well-recognized r_j and/or missed and thus, unreliable elements $u_j = \overline{r_j}$, and the probability of occurrence of such an event is calculated from the probabilities of occurrence of the elements. Following [1], where it is implicitly assumed that the elements are statistically independent, a recognition event Q_i can be expressed by the product of correct element recognition probabilities p_i's with $i \in r_j$ and element error probabilities $(1-p_i)$'s with $i \in u_j$, describing this event.

The parameter Q_i signifies, in our model, the probability that i time-frequency blocks were correctly recognized (by i recognizers, each of them only using the information from the element at its input, i.e. only 'sensory' information) and that $(n-i)$ blocks were missed. This includes for each Q_i all permutations of i elements, so that we can write for a multi-band system of n subbands (with $p_i = P(q_k|x_i)$ for x_i the i^{th} frequency element):

$$
\begin{aligned}
Q_0 &= P(q_k|x_1)P(q_k|x_2)...P(q_k|x_n), \\
Q_1 &= (1 - P(q_k|x_1))P(q_k|x_2)...P(q_k|x_n) + ... + \\
&\quad P(q_k|x_1)...P(q_k|x_{n-1})(1 - P(q_k|x_n)) \\
&\vdots \\
Q_n &= (1 - P(q_k|x_1))(1 - P(q_k|x_2))...(1 - P(q_k|x_n))
\end{aligned} \tag{3}
$$

2.2.2 Measure of contextual information

Let us now come to the question of how to model the context parameters c_i in the case of multi-band ASR using n frequency subbands. If an automatic recognizer is "asked" to make a guess at an element it previously mis-classified, the probability of correct classification would remain equal to the phoneme prior probability. This would be the same for each missed element, thus, $c_i = c = P(q_k)$.

Following (1), we now multiply the probability of occurrence of each recognition event Q_i by the chance c^i of guessing the $i = |u_j|$ missed parts, and explicitly sum over *all possible events* $j = 1..N$ ($n =$ number of frequency subbands and $N = 2^n$), obtaining the recognition probability of the combined multi-band system of n subbands for phoneme q_k as

$$
\begin{aligned}
P(q_k|x) &\simeq \sum_{i=0}^{n} c^i Q_i \\
&\simeq \sum_{j=0}^{N} \prod_{i \in r_j} P(q_k|x_i) \prod_{l \in u_j} (1-P(q_k|x_l))\, c^{|u_j|}
\end{aligned} \tag{4}
$$

with r_j denoting the correctly recognized (reliable) elements and u_j the erroneously recognized (unreliable) elements of event j ($j = 1..N$). For a certain combination j, the weight amounts to $c^{|u_j|}$ with $|u_j|$ the number of mis-classified elements in combination j.

3 FC Subband Processing

In our earlier work on multi-band processing, we proposed the FC model which also considers all possible combinations of (reliable) subband classifiers, combining them in a weighted sum [5]. One strategy to implement such a system in the framework of HMM/MLP hybrid systems, is by training a separate MLP expert on each combination j of subbands, which results in the **FC formula** :

$$\begin{aligned} P(q_k|x) &= \sum_{j=0}^{N} P(r_j|x)P(q_k|r_j,x) \\ &\simeq \sum_{j=0}^{N} w_j \, P(q_k|x_{r_j}) \end{aligned} \tag{5}$$

with w_j the reliability weight for expert j, and x_{r_j} the combination j of subbands $i \in r_j$ ($x_{r_j} = \cup_{i\in r_j} x_i$).

As such a system demands a lot of training and parameters, an approximation (AFC) was developed [3] (assuming conditional independence of the one-band streams) which only employs the MLP experts trained on the one-band streams, the outputs of which are then used during recognition to approximate all other combination probabilities $P_{r_j}(q_k|x)$ as follows:

$$P_{r_j}(q_k|x) \simeq \Theta_k P^{1-|r_j|}(q_k) \prod_{i\in r_j} P(q_k|x_i) \tag{6}$$

with Θ_k a normalization factor. We then write for the **AFC formula** :

$$P(q_k|x) \simeq \sum_{j=0}^{N} w_j \, \Theta_k P^{1-|r_j|}(q_k) \prod_{i\in r_j} P(q_k|x_i) \tag{7}$$

$$\simeq \sum_{j=0}^{N} w_j \, \frac{P_{r_j}(q_k|x)}{\sum_{k'=1}^{K} P_{r_j}(q_{k'}|x)} \tag{8}$$

with $w_j = P(r_j|x)$ reliability weight for approximation j. For correct approximation of the combination probabilities by the single-stream probabilities, the factor $P^{1-|r_j|}(q_k)$ is essential. In (8) each approximated combination $P_{r_j}(q_k|x)$ is then normalized over the set of all phonemes k ($k = 1..K$). For derivation of this approach, see [3, 2].

4 Comparing AFC and FC to the new model for error correction

In this section, we compare the new formalism (4) which we obtained from the interpretation of Bronkhorst et al.'s model to the two multi-band approaches AFC and FC. We will see how (4) can be used to extend, first, the AFC model, which bases on a similar assumption of independence between the element probabilities. We can then apply this extension, which takes account of the error probabilities from the unreliable bands together with a correction factor, also to the FC model. In this case, the assumption of independence is no longer needed.

4.1 AFC with Error Correction (AFC-ECPC)

Comparing the AFC formula (7) to the recombination formula which we obtained by application of Bronkhorst et al.'s model (4), we can see the similarity between the two approaches. The main differences are that in Bronkhorst et al.'s model the error probabilities of the supposedly unreliable bands are included by the second term in (4) and that in the AFC model the factors $P^{1-|r_j|}(q_k)$ and Θ_k have to be considered. Moreover, the weighting factors are interpreted differently: in (4), the weights c_i are interpreted as context information which is used to correct the wrongly identified elements $u_j = \overline{r_j}$, whereas in our former system (7) the weights w_j indicated the reliability of the correctly recognized elements r_j.

We therefore decided to combine both approaches in a joint model. This can be realized by extending the AFC approach to also admit the respective error probability in the calculation of each combination probability and by adapting the interpretation of the weights. The new AFC approach with "error correction in posterior combination" **AFC-ECPC formula** simply results in:

$$\begin{aligned} P(q_k|x) &\simeq \sum_{j=0}^{N} \Theta_k \frac{\prod_{i\in r_j} P(q_k|x_i)}{P^{|r_j|-1}(q_k)} \prod_{l\in u_j} (1-P(q_k|x_l))\, c^{|u_j|} \\ &\simeq \sum_{j=0}^{N} \frac{P_{r_j}(q_k|x)}{\sum_{k'=1}^{K} P_{r_j}(q_{k'}|x)} \prod_{l\in u_j} (1-P(q_k|x_l))\, c^{|u_j|} \end{aligned} \tag{9}$$

with $i\in r_j$ denoting the correctly recognized elements i, $l\in u_j$ the wrongly identified elements l, and $P_{r_j}(q_k|x)$ as in (6).

4.2 FC with Error Correction (FC-ECPC)

To overcome the assumption of independence between the elements, which is for an automatic speech recognizer a strong restriction, we substitute the approximated combinations by one recognizer each, returning to the FC approach. As we saw above, the application of Bronkhorst et al.'s model resulted in a context model for the AFC approach which we denoted as AFC-ECPC. If we now apply it to FC, we obtain the **FC-ECPC formula**[2] as:

$$P(q_k|x) \quad = \quad \sum_{j=0}^{N} P(q_k|x_{r_j})(1-P(q_k|x_{u_j}))\, c^{|u_j|} \tag{10}$$

For each position of reliable data x_{r_j} $(j=1..N)$, the error of the unreliable part x_{u_j} is not discarded but, each time, multiplied as an error probability to the posterior probability of the reliable data. This error is then accounted for through the multiplication by the respective power of $c = P(q_k)$ (which is an initial, very rough model for the context information as described in Section 2.2.2 above).

5 Experiments

Experiments were carried out on a test set of 200 utterances from the Numbers95 database of connected numbers recorded over the telephone line. For tests on noise-corrupted data, real-environmental car noise (from an in-house database of car noise by Daimler Chrysler) and factory noise (from the Noisex92 database) were added at signal-to-noise ratio (SNR) values of 12 and 0 dB to the clean test data.

The HMM/MLP hybrid systems, using j-rasta features, were trained on the (clean) Numbers95 training part. Our multi-band system consists of 4 frequency subbands, so that the FC and FC-ECPC systems comprise 16 MLP experts (trained on each combination of subbands), whereas the AFC and AFC-ECPC systems only use the 4 MLPs trained on one frequency subband each.

5.1 Experiments with AFC-ECPC

As the application of Bronkhorst et al.'s model to ASR rather resembles the AFC than the FC approach, we started by testing the former (although this model usually results in lower recognition

[2]ECPC=Error Correction in Posterior Combination.

Rule	Car 0 dB	Car 12 dB	Factory 0 dB	Factory 12 dB	Clean 45 dB
AFC equal weights	47.0	16.4	46.6	16.9	12.5
AFC proposed wghts	48.0	17.2	47.0	17.2	12.1
AFC-ECPC equal weights	46.1	16.1	46.0	16.8	12.2
AFC-ECPC priors as wghts	47.1	17.9	51.1	17.9	13.9

Table 1: *Word error rates (WER) for the approximated full combination (AFC) system and its extended version (AFC-ECPC) using error correction as proposed by [1], conducted on clean data and two noise cases (car and factory noise) at 0 and 12 dB SNR. For description of the weights see text.*

rates except for band-limited noise [3]). In order to see the difference resulting from including the error probabilities, we first tested both systems on the same weights which is only possible if the weights are chosen as $c = P(r_j|x) = c^{|u_j|} = 1$. Results in Table 1 ($1^{st}$ and 3^{rd} line) show that both systems result in similar performance in clean and noise, with a slight gain for AFC-ECPC.

In the next set of experiments, each system was tested with its proposed weights, i.e. $P(r_j|x) = 2^{|r_j|}$ for AFC (which is increasing for the number of (reliable) bands in an approximated combination, thus for each subband in a combination we simply multiplied by the factor 2) and $c^{|u_j|} = P^{|u_j|}(q_k) = 0.037^{|u_j|}$ for AFC-ECPC assuming equal priors of the 27 phonemes in the database (2^{nd} and 4^{th} line in Table 1). Here, both systems suffered a slight deterioration (except for AFC on clean). (In former experiments, we had also found that AFC performs best on equal weights.)

5.2 Experiments with FC-ECPC

In order to release the assumption of independence among frequency subbands, we turn to the FC systems. Again, we first tested the original FC system and the extended version FC-ECPC setting the weights to $c = 1$. No significant difference in the results of both systems were observed (cf. Table 2, 1^{st} and 3^{rd} line).

Rule	Car 0 dB	Car 12 dB	Factory 0 dB	Factory 12 dB	Clean 45 dB
FC equal weights	39.9	10.6	40.9	11.8	8.6
FC RF weights	35.1	10.1	36.8	10.0	8.0
FC-ECPC equal weights	39.0	10.8	41.1	11.9	8.4
FC-ECPC priors as wghts	**31.2**	11.1	**33.8**	12.0	8.6

Table 2: *WER for the full combination (FC) system and its extended version (FC-ECPC) using error correction as proposed by [1], conducted on clean data and two noise cases (car and factory noise) at 0 and 12 dB SNR. For description of the weights see text.*

In the next set of experiments, we chose for the FC system our (so far) best non-adaptive weights, which are relative frequency (RF) weights calculated on the training data [4]. This was compared to FC-ECPC using error correcting weights as described above (cf. Table 2, 2^{nd} and 4^{th} line). Both systems improved significantly in high levels of noise when the appropriate weights were used, whereas performance stayed (almost) the same for low level noise and clean speech. The new FC-ECPC system resulted in especially good performance at 0 dB SNR. Therefore, we conclude that the gain of the ECPC system does not stem alone from the inclusion of the error probabilities from the non-reliable bands, but to a high degree depends on the correct choice of error correction by an appropriate selection of the correction factor $c^{|u_j|}$.

6 Conclusion

In this article, we proposed a new approach to multi-band ASR which was derived from a "model for context effects in human speech recognition" proposed by [1]. The similarities between the new models, FC-ECPC and AFC-ECPC, and former multi-band systems were discussed. The performance of the error correction mechanism in the new models for automatic speech recognition depends to a large extend on the appropriate choice of weights. An initial interpretation of the weights was proposed and showed very promising results for FC-ECPC in noise corrupted speech. Just as in the case of AFC, the approximated system, AFC-ECPC, cannot compete with the FC-ECPC system when tested on wide-band noise [3]. We are planning on investigating other weighting strategies to improve the performance of the new error correcting multi-band models also in higher SNR and clean speech.

References

[1] A.W. Bronkhorst, A.J. Bosman, and G.F. Smoorenburg. A model for context effects in speech recognition. *Journal of the Acoustic Society of America*, 93(1):499–509, 1993.

[2] A. Hagen and H. Glotin. Études comparatives des robustesses au bruit de l'approche 'full combination' et de son approximation. *Journée d'Études sur la Parole, Aussois*, pages 317–320, 2000.

[3] A. Hagen, A. Morris, and H. Bourlard. Subband-based speech recognition in noisy conditions: The full combination approach. IDIAP-RR 15, IDIAP, 1998.

[4] A. Hagen, A. Morris, and H. Bourlard. From multi-band full combination to multi-stream full combination processing in robust ASR. *ISCA ITRW ASR2000*, pages 175–180, 2000.

[5] A. Morris, A. Hagen, H. Glotin, and H. Bourlard. Multi-stream adaptive evidence combination to noise robust ASR. *Speech Communication*, 34(1-2), 2001.