Performance of Instantaneous and Time-Difference Features in "Full Combination" Multi-Stream processing for Robust ASR

Astrid Hagen
Spoken Language Systems Lab, INESC-ID, Rua Alves Redol 9, Lisbon, Portugal.
astrid.hagen@l2f.inesc-id.pt
formerly at: IDIAP, Rue du Simplon 4, Martigny, Switzerland,
where this work was carried out.

Abstract

A means to achieve higher noise robustness of an automatic speech recognizer is, amongst others, multi-stream processing, where the independent errors of the streams can usually be dampened in the combination process, so that the final system achieves higher recognition rates in any environment. A crucial question in multi-stream processing is how the necessary diversity in the streams can be achieved. In this article, we discuss the multi-stream approach of using diverse feature streams, more specifically, combinations of short- and long-term features. In such a setup, it usually needs to be decided whether streams are recombined before or after acoustic modeling. We show how both approaches are combined by the use of "full combination" multi-stream processing and demonstrate its increased performance in several acoustic environments as compared to standard (multi- and one-stream) processing.

1 Motivation

In Multi-Stream (MS) processing several input streams consisting in *different representations of the same source* are processed in parallel instead of only processing one stream as usually done in standard processing. MS recognition is, amongst others, based on the observation that different representations of the speech signal often lead to different kinds of recognition errors or the same errors occurring at different points. The different streams are thus expected to complement each other at the recombination stage and to lead to a more powerful and robust performance of the combined MS system.

The paradigm of using an ensemble of trained classifiers instead of simply using only one classifier has been widely proposed in the literature (Jordan and Jacobs, 1994; Bishop, 1995). The idea behind using multiple classifiers is that, in the absence of the "true" model, the apparent best classifier can be improved upon by employing several classifiers to solve the classification task independently and then construct a final decision by making use of the individual scores (Hashem, 1997).

MS processing is also motivated from psycho-acoustic research where it was found that *multiple representations* of the speech signal and *appropriate time-scales* are employed in human auditory processing to render the final representation as robust as possible (Yang et al., 1992).

With the wide variety of *feature extractors* available for automatic speech recognition (ASR), a promising approach thus seems to be to base the diversity of ensemble classifiers on the diversity of the input streams. Acoustic features are often known to work especially well under certain environmental conditions, whereas others show their advantage under completely different conditions.

The use of several feature streams not only comprises employing different *kinds* of features (as e.g. PLP and MFCC features (Hagen et al., 2000)) but also different processing strategies for the same kind of features, such as extraction from different time scales or employing various pre- or post-processing strategies. Here, the shorter time scales are often found to provide good performance in clean speech whereas the longer time scales usually show higher noise robustness (Sharma et al., 2000; Weber, 2000).

Larger time scale features which are often used to complement instantaneous (RAW) features in state-of-the-art ASR are the temporal first- and second-order derivative features (also refered to as DELTA and DDELTA features, respectively). Besides being independent of the raw features, they have the advantage that they can be easily derived from the instantaneous features without the need of additional feature extraction. We propose in this article to use these features as separate streams in the MS framework, more specifically, in "full combination" MS processing, which is described next.

2 "Full Combination" Multi-Stream

2.1 The Multi-Stream Paradigm

Given several different feature streams, we can distinguish in MS processing between feature combination and probability combination.

In *feature combination*, the feature vectors are combined before acoustic modeling. This is what is usually done in most of the state-of-the-art ASR

systems, where for example the first- and second-order time or frequency derivative features are concatenated to the instantaneous features.

In *probability combination*, the different streams are processed by specific acoustic models, the outputs of which are then recombined. After processing the different streams for feature extraction and probability estimation (cf. **Figure 1**), different probability combination strategies, such as the sum or product rule, can be applied.

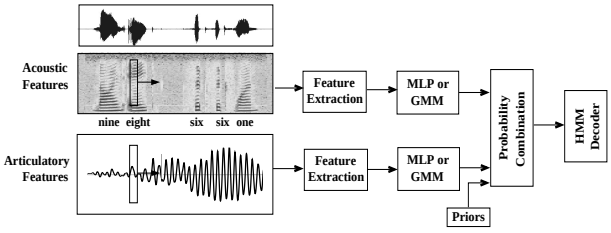

Figure 1: Illustration of probability combination in MS ASR on two streams using different feature sets.

It cannot be known without extensive testing whether feature or probability combination or some combination of both is best suited for a given task. We therefore propose to follow the "full combination (FC)" scheme as introduced for Multi-Band (MB) processing (Morris et al., 2001; Hagen, 2001) also in MS processing, in which both approaches are sensibly combined.

2.2 Full Combination Processing

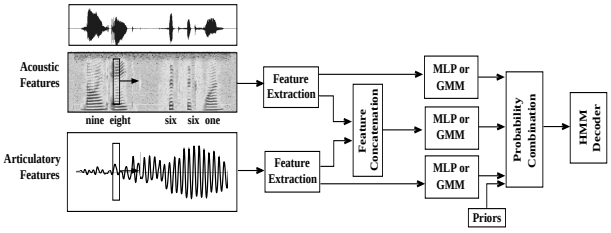

Figure 2: Illustration of "full combination" in MS ASR on two streams using different feature sets.

It is supposed that at each instant one combination of data streams carries the clean data and is best suited for identifying the current phoneme[1], as one or the other combination has to be best (Hagen, 2001). As it is not known which combination of streams comprises clean data features it has to be integrated over all $2^S = \mathcal{B}$ possible combinations s_i $(i = 1, \ldots, \mathcal{B})$, with S the number of individual feature streams.

In order to integrate over all possible combinations of streams, there are two steps involved in FC MS processing (Morris et al., 2001; Hagen, 2001), which can be seen in **Figure 2**.

- First, the feature vectors from all streams have to be concatenated into all possible combinations of feature vectors.
- Second, each stream combination is processed independently by different experts.

The probability estimates from the different experts are then recombined according to the FC formulae described in the following.

As we work with Hidden Markov Model/Multi-Layer Perceptron (HMM/MLP) hybrid systems, we present the FC recombination formulae for posterior-based systems only. Using s_i with $i = 1, \ldots, \mathcal{B}$ to denote a certain stream at time t, the following rules can be established (Hagen, 2001). The FC SUM rule:

$$P(q_k|s) = \sum_{i=1}^{\mathcal{B}} P(q_k|s_i)P(b_i|s) \quad (1)$$

with $P(q_k|s_i)$ the probability estimate for phoneme q_k $(k = 1, \ldots, K)$ from expert i trained on stream s_i, and $P(b_i|s)$ the reliability term for expert i given acoustic vector s; the FC PRODUCT rule:

$$P(q_k|s) = \Theta_k \, \Theta \, \frac{\prod_{i=1}^{\mathcal{B}} P^{w_i}(q_k|s_i)}{P^{(\sum_i w_i)-1}(q_k)} \quad (2)$$

with $P(q_k)$ the class priors and w_i the stream weights. The FC PRODUCT WITH EQUAL PRIORS :

$$P(q_k|s) = \Theta \prod_{i=1}^{\mathcal{B}} P^{w_i}(q_k|s_i) \quad (3)$$

with Θ_k a normalization constant dependent of k, and Θ a normalization constant independent of k, such that $\sum_{k=1}^{K} P(q_k|s) = 1$ (Hagen, 2001, p.86f). Equation (3) assumes equal class priors which is often used to approximate the product rule.

3 Experiments and Results

In our multiple-time-scale MS system, the three proposed combination algorithms, employing equal weighting, were tested on a continuously spoken digits database (Numbers95) (Cole et al., 1995) under noise-free (matched) conditions and under noise-corruption (mismatch) by band-limited (stationary and siren) and wideband noise (Hagen, 2001). The artificially created stationary band-limited noise comprises 4 noise cases (in different frequency bands), the real-environmental wideband noise comprises car and factory noise, each at 0 and 12 dB signal-to-noise ratio (SNR). The word error rates (WERs) presented here were calculated as average values over the respective noise cases. All

[1] This is under the assumption that the stream acoustic models are trained on clean speech only.

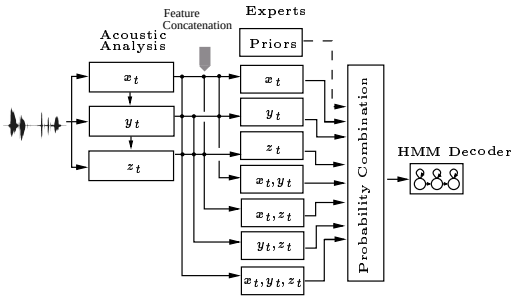

Figure 3: Illustration of FC recognizers combination, using RAW (x_t), DELTA (y_t) and DDELTA (z_t) features as individual input streams as well as all possible combinations of feature streams in the framework of an HMM/MLP system.

tests were run using both PLP and J-RASTA-PLP features. The 12 RAW features (together with the energy) were calculated on windows of 25 ms length, with a shift of 12.5 ms.

For each feature set, the first-order difference features are calculated over five raw features ($y_t = [-2x_{t-2} - x_{t-1} + x_{t+1} + 2x_{t+2}]$), and the second-order difference features over seven raw features ($z_t = [2x_{t-3} + x_{t-2} - 2x_{t-1} - 2x_t - 2x_{t+1} + x_{t+2} + 2x_{t+3}]$). They cover respectively 75 ms and 100 ms of speech data. For each of the seven streams[2] (illustrated in **Figure 3**) an MLP recognizer was trained, the number of parameters of which vary between 144 000 and 661 500, depending on the size of the input stream and the hidden layer. The posterior probabilities at the output of the MLPs are combined via FC SUM (1), FC PRODUCT (2) and FC PRODUCT WITH EQUAL PRIORS (3), using equal weights.

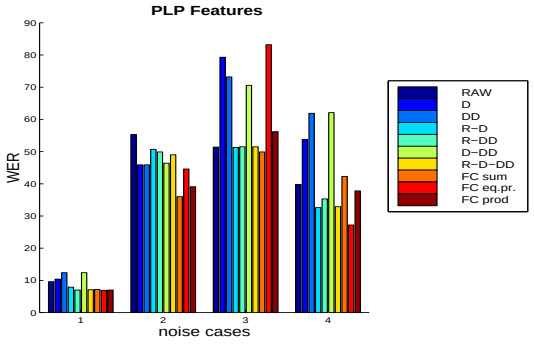

Figure 4: WERs of each constituent stream tested by itself and after FC combination for clean speech and speech corrupted with various additive noise cases.

The results of the **PLP** baseline system (7^{th} bar, referred to as RAW-D-DD[3]), together with the recognition performance of each of the other six constituent streams, as well as the performance after FC probability combination are shown in **Figure 4** for clean speech and the different noise cases.

In clean speech (case 1), none of the three combination schemes (last 3 bars in each case in Figure 4) performs significantly different from the baseline. When looking at the performance of the constituent streams, we see that all combinations including the RAW features perform equally well as the baseline. Only the single streams and the combination of DELTA and DDELTA features deteriorate (D-DD) when used by themselves. Similar results were found in (Macho et al., 1999, p. 113) on isolated digit recognition where the "acceleration filter" (i.e. the DDELTA features) also hurt performance in clean speech. The authors argue that degradation in clean speech of the DDELTA features could be due to the complete cancellation of the zeroth modulation frequency. Thus, static features are needed to achieve good performance in clean speech.

In stationary band-limited noise (case 2), FC SUM and FC PRODUCT perform significantly better than all other streams. The FC PRODUCT WITH EQUAL PRIORS results in a weaker (though significant) improvement. It is interesting to note that, in these noise conditions, the DELTA and DDELTA streams as well as their combination outperform any other constituent stream and the baseline.

In non-stationary band-limited noise (case 3), these three streams deteriorate the most (beside the FC PRODUCT (WITH EQUAL PRIORS)), whereas all streams which include the static features are more noise robust. Here, no significant performance difference is observed between the RAW-D-DD baseline and the FC SUM, though the FC PRODUCT and FC PRODUCT WITH EQUAL PRIORS deteriorate significantly.

The experiments on real-environmental noise (case 4) led to rather different results than for the other noises. Here, it is FC PRODUCT WITH EQUAL PRIORS which outperforms all other streams. Simple feature concatenation of the RAW features with one of the other two streams also results in performance competitive to the baseline.

J-RASTA Features We now turn to using J-RASTA-PLP features. As we have often observed in our experiments, J-RASTA-PLP -based systems are harder to improve as their overall performance is already higher.

In clean speech (case 1 of **Figure 5**), only the DELTA and DDELTA feature streams (used by themselves) led to degradation for the J-RASTA-PLP features, whereas for PLP features, also the RAW stream by itself was not competitive. All other streams and combination schemes led to similar performance.

[2]This disregards the empty stream.

[3]This refers to the concatenated features used at the input to the stream MLP : "RAW" denotes the static (RAW) features, "D" the DELTA features, and "DD" the DDELTA features, as usually concatenated, thus "baseline".

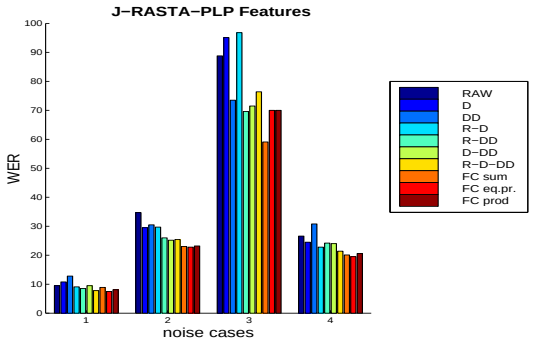

Figure 5: WERs of each constituent stream tested by itself and after FC combination for clean speech and speech corrupted with various additive noise cases.

In stationary band-limited noise (case 2), we can again observe that the RAW features by themselves degrade most, though less than for PLP features. All three FC strategies perform equally well in these conditions and better than any other system. In case of non-stationary narrow-band noise (case 3), the first-order derivative features give among the worst performance for both feature sets, whereas the second-order derivatives degrade significantly less. Combinations with this feature stream even lead to the most robust concatenated feature streams (RAW-DDELTA and DELTA-DDELTA) in the case of J-RASTA-PLP processing, (whereas for PLP features it was the combinations with the RAW features). FC performs at least as well and significantly better when the FC SUM was employed. In wide-band noise (case 4), it is the DDELTAS which cannot handle this noise corruption. Here, the RAW features are needed in both feature sets to enhance performance of the constituent streams. FC processing again achieves among the best results.

As it could be seen, the relative improvement is in general less when employing J-RASTA-PLP features. This can be due to the fact that (J-)RASTA processing is very similar to the calculation of the difference features and that less gain is achieved when both are used jointly.

The above results show that each of the three feature streams (RAW, DELTA and DDELTA) is powerful on a different kind of noise condition. By FC processing we are able to better exploit this characteristic which can be seen through the enhanced recognition performance of at least one of the FC systems over the baseline setup for each noise case, especially in the case of J-RASTA-PLP features.

4 Conclusion

To sum up, in clean speech, FC processing did not result in any significant improvement as compared to pure feature concatenation. In stationary band-limited noise, FC SUM and FC PRODUCT were significantly better than the baseline for PLP features, and slightly better for J-RASTA-PLP. In non-stationary noise, the FC SUM gave improved performance. In real-environmental noise, the FC PRODUCT WITH EQUAL PRIORS using PLP features significantly improved over the baseline. With the RAW, DELTA and DDELTA features showing very different performance amongst each other and in the different noise cases, simple feature concatenation does not account for their individual potential. Thus, employing them in a FC MS setup allows us in a better way to exploit the advantage of each feature set and combination of sets, and led to better performance in most acoustic environments which were tested. The problem which remains is how to decide which FC combination strategy is to be employed for unseen noise.

References

Ch.M. Bishop. 1995. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

R.A. Cole, M. Noel, T. Lander, and T. Durham. 1995. New telephone speech corpora at CSLU. *EUROSPEECH*, 1:821–824.

A. Hagen, A. Morris, and H. Bourlard. 2000. From multi-band full combination to multi-stream full combination processing in robust ASR. *ASRU*, pages 175–180.

Astrid Hagen. 2001. *Robust speech recognition based on multi-stream processing*. Ph.D. thesis, École Polytechnique Fédérale de Lausanne, Switzerland.

S. Hashem. 1997. Optimal linear combination of neural networks. *Neural Networks*, 10(4):599–614.

M. I. Jordan and R. A. Jacobs. 1994. Hierarchical mixture of experts and the EM algorithm. *Neural Computation*, 6:181–214.

D. Macho, C. Nadeu, J. Hernando, and J. Padrell. 1999. Time and frequency filtering for speech recognition in real noise conditions. *ESCA Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pages 111–114.

A. Morris, A. Hagen, H. Glotin, and H. Bourlard. 2001. Multi-stream adaptive evidence combination to noise robust ASR. *Speech Communication*, 34(1-2):25–40.

A. Sharma, D. Ellis, S. Kajarekar, P. Jain, and H. Hermansky. 2000. Feature extraction using non-linear transformation for robust speech recognition on the aurora database. *ICASSP*, 2:1117–1120.

K. Weber. 2000. Multiple time scale feature combination towards robust speech recognition. *Konvens*, pages 295–299.

X. Yang, K. Wang, and S. A. Shamma. 1992. Auditory representations of acoustic signals. *IEEE Trans. on Inf. Theory*, 38(2):824–839.