MULTI-STREAM PROCESSING USING CONTEXT-INDEPENDENT AND CONTEXT-DEPENDENT HYBRID SYSTEMS

Astrid Hagen[1], *João P. Neto*[1,2]

[1]L[2]F – Spoken Language Systems Laboratory, INESC-ID Lisbon, Portugal
[2]Instituto Superior Técnico, Portugal

{Astrid.Hagen,Joao.Neto}@l2f.inesc-id.pt

ABSTRACT

Multi-stream processing provides a successful approach to enhance the generalization capability of a recognizer and can, moreover, be combined with other robust techniques, such as spectral subtraction and/or robust features, HMM/MLP hybrid systems, and others. The question usually arises at which point the different streams are to be recombined, i.e. at the feature or at the probability level. Feature and probability combination are often seen as alternative approaches. We show here how a sensitive combination of both renders this decision obsolete and improves recognition as compared to each approach carried out on its own.

The study has been carried out on the digits and numbers part of the Portuguese SPEECHDAT corpus. This corpus includes a large number of speakers and channel conditions and is, thus, well suited to test the described multi-stream systems under realistic conditions. Results are presented for both context-independent and context-dependent models used in an HMM/MLP hybrid system.

1. INTRODUCTION

Many speech recognizers are applied over the telephone line and employ digit and number recognition. This includes such applications as credit card and account number validation, automated dialing, user identification via PIN codes, and others. In these tasks, the speech recognizer is confronted not only with a large number of speakers with different characteristics but also with a wide variety of transmission channels which alter the speech signal. The SPEECHDAT database was developed with the goal to provide the research community (and industry) with a realistic speech corpus to test and develop speech recognition tools which can account for such real-world applications.

In these environments, state-of-the-art recognizers, such as HMM/MLP (Hidden Markov Model/Multi-Layer Perceptron) hybrid systems, together with robust features, e.g. RASTA or MFCC features possibly with spectral subtraction or other additional filters, usually provide some robustness. A successful approach to further enhance the performance of such a speech recognizer to unseen conditions is multi-stream (MS) processing.

In MS processing several information streams are processed in parallel up to a certain point where the information is recombined to obtain one final decision. It was found that the more diverse the streams are, the better they complement each other and the higher the gain in recognition rate usually is. The different streams can consist of (i) different modalities (e.g. audio and video data), (ii) different acoustic models (AMs), training data and/or algorithms, or (iii) different feature streams.

In case when the diversity of the streams is already obtained before the AMs, recombination can be carried out at two distinct levels: the feature level (i.e. before the AMs) [1, 2], and the probability level (i.e. after the AMs and before or during decoding) [3, 4, 5]. These approaches are usually interpreted as alternatives, hypothesizing that correlated features should be modeled jointly, whereas uncorrelated features should be modeled by disjoint acoustic models. This however, could not be sustained by experiments [2]. More recently, the "Full Combination" (FC) approach was proposed which is a mathematically correct extension of standard probability combination and actually combines both methods [3, 6]. Higher recognition rates are usually achieved through FC processing than with either method, feature or probability combination, on its own.

In this article, we investigate the MS FC approach on the Portuguese SPEECHDAT database. Our streams stem from three different, state-of-the-art acoustic feature extractors which are known to be powerful in rather diverse conditions and thus complement each other well. In the next section, we give the mathematical background for the probability combination approaches which were tested. In Section 3, the SPEECHDAT database and the experiments are described for context-independent (CI) and context-dependent (CD) models. In the last section we summarize the results and describe our ideas for future work.

2. PROBABILITY COMBINATION STRATEGIES

Non-linear recombination by product rule is one of the most widely used combination strategies for probability estimates. This rule assumes independence of the (in our case posterior) probabilities of one class given the data from different streams, which amounts to assuming equal class priors.

$$P(q_k|x) \quad = \quad \Theta \prod_{i=1}^{S} P_i(q_k|x) \tag{1}$$

where S is the number of individual streams and $P_i(q_k|x)$ the probability estimate for speech unit q_k $(k = 1, \ldots, K)$ from ex-

pert i which is trained on input data x. Θ is a normalization constant, independent of q_k, such that $\sum_k P(q_k|x)=1$.

The (standard) sum rule for posteriors is written as follows:

$$P(q_k|x) = \sum_{i=1}^{S} P(q_k|b_i,x)P(b_i|x) \quad (2)$$

$$\simeq \sum_{i=1}^{S} P_i(q_k|x)P(b_i|x) \quad (3)$$

where S and $P_i(q_k|x)$ as above. $P(b_i|x)$ is a weighting term which depends on both the expert i and the acoustic vector x. The weights can be calculated e.g. offline on the training data via Least-Mean Square Error estimation or online during decoding via Signal-To-Noise Ratio estimation [3, 6].

In the above approach, the set of experts is not exhaustive and it can happen that the best combination of streams is simply ignored. The FC approach, on the contrary, considers all possible combinations of streams by defining a set of exhaustive and mutually exclusive experts (cf. Figure 1). As it is not known which combination of streams comprises the best data features for the current frame, it has to be integrated over all $2^S = \mathcal{B}$ possible combinations x_j $(j = 1,\ldots,\mathcal{B})$, with S the number of individual feature streams. This amounts to the FC SUM rule:

$$P(q_k|x) = \sum_{j=1}^{\mathcal{B}} P_j(q_k|b_j,x)P(b_j|x) \quad (4)$$

$$= \sum_{j=1}^{\mathcal{B}} P_j(q_k|x)P(b_j|x) \quad (5)$$

with $P_j(q_k|x)$ the probability estimate for phoneme q_k from expert j trained on its stream (combination) x, and $P(b_j|x)$ the weight for expert j given acoustic vector x.

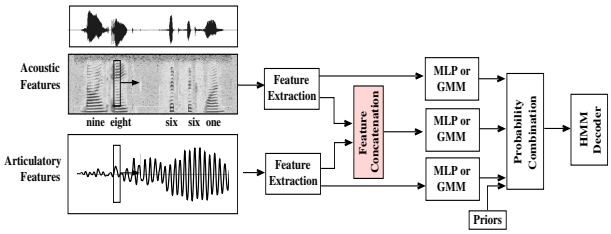

Fig. 1. Illustration of "Full Combination" processing in the multi-stream approach on two streams using different feature sets.

We will see in the experiments how the FC approach leads to consistently better performance than either pure feature concatenation or the simple probability combination strategies on their own.

So far, FC processing has only been applied to context-independent models [3, 6]. We will discuss in the next paragraph the advantages of context-dependent processing for which the above combination strategies also apply without any changes.

Context-Dependent Modeling

In standard HMM/GMMs (Gaussian Mixture Models) the use of context-dependent models significantly improves recognition accuracy due to the explicit modeling of coarticulation effects producing sharper probability density functions for the different phone classes. We work in the framework of HMM/MLP hybrid systems where the posterior probabilities at the output of the MLP are, after division by the priors, used as scaled likelihoods in the HMM for decoding. In hybrid systems, usually only one state is used per phone model together with duration-modeling. For this reason, the modeling capabilities of the hybrid systems are limited, especially as far as detailed modeling of the phone changes is concerned. To circumvent this problem, these systems use a context window at the input of the MLP. In order to combine the advantages of discriminative training with the advantages of context-dependent modeling we introduce triphone models to our hybrid systems. The use of triphones implies enlarging the output layer of the MLP. More (speech unit) classes at the output of the MLP renders the MLP more difficult to train and increases the need for more training data. For the digits and numbers task, this is still feasible as the number of occurring triphones is limited and the size of the MLP's output layer will not increase too much.

3. EXPERIMENTS

The Portuguese SPEECHDAT database has been developed within the SPEECHDAT project[1] to address current and future requirements in the field of telecommunication, spoken language technology and research. It has been recorded in two phases over the public telephone network involving a large set of speakers, recording conditions and tasks. In the first phase (SPEECHDAT 1), there were 1000 speakers. Thirteen different noise cases were manually marked (start and end point) on the speech data. In the second phase (SPEECHDAT 2), 4000 speakers were involved. The noise cases were merged into 4 remaining classes and roughly marked in every utterance.

In this article we concentrate on the digits and numbers part of the database, more precisely the categories B1, C1–4 and I1 described in Table 1.

The training and cross-validation set comprises 9981 clean[2] utterances (13h 24min of speech), roughly equally distributed in terms of utterances over the six numbers categories as shown in the left table of Figure 2. The test set consists of 929 clean utterances (1h 14min of speech), distributed as shown in the right table of Figure 2. The sets correspond to the defined partitioning of the speakers into training and test set as given on the SPEECHDAT CDs, so that each speaker was only used in either of the sets.

An alignment was created with Gaussian models, using flat start, and then refined with the MLPs, using the clean utterances of the better labeled first part of the corpus (SPEECHDAT 1). These MLPs were then used to align the clean utterances of the second part (SPEECHDAT 2). The whole set of training utterances was then re-aligned several times.

In this work we use 3 feature streams comprising 13 PLP cepstra, 13 RASTA(-PLP) cepstra and 28 Modulation Spectrogram (MSG) features, extracted on windows of 20ms with a frame shift of 10ms. The first two streams were augmented by their delta features.

The MLP uses 7 frames of context information except for the MSG features where 9 frames are used. The hidden layer consists

[1] `http://www.speechdat.org`

[2] "Clean" here signifies no speaker or background noise though moderate noise introduced by the telephone network is a natural consequence of the recording conditions.

Class ID	Class contents	Example To read	As has been read
B1	10 isolated digits	0965423871	”zero nove seis cinco quatro dois três oito sete um”
C1	Sheet number	33546	”três três cinco quatro seis”
C2	Telephone number	090981696	”zero noventa nove oito um seis nove seis”
C3	Credit card number	4585 4567 6189 6565	”quatro mil quinhentos e oitenta e cinco quatro mil ...”
C4	PIN code	159.160	”cento e cinquenta e nove mil cento e sessenta”
I1	1 isolated digit	6	”seis”

Table 1. Illustration of the digits and numbers classes of the SPEECHDAT database. As can be seen in some of the examples, some of the digits have actually been read as connected numbers (e.g. ”noventa”).

Training set	
B1:	1461
C1:	1606
C2:	1770
C3:	1621
C4:	1566
I1:	1957
SUM	9981

Test set	
B1:	110
C1:	144
C2:	179
C3:	180
C4:	117
I1:	199
SUM	929

Fig. 2. Distribution of the utterances in the training and cross-validation set (left) and in the test set (right) over the six classes of the SPEECHDAT database used here.

of 2000 nodes (2770 for MSG), and the number of output nodes corresponds to the number of speech units in the digits and numbers part of the SPEECHDAT corpus.

The vocabulary consists of 51 words for which an internal transcription was available. The language model (LM) was set up on the training utterances, using the CMU-Cambridge Language Modeling Toolkit V2.05. The Good-Turing method was used to estimate the closed-vocabulary, back-off bigram LM which contains 2601 bigrams. Missing bigram combinations which did not occur in the training data were manually added. The perplexity of the LM on the test set is 10.73.

The hybrid systems employing context-independent (monophone) models use 32 MLP output nodes (one for silence) as only 31 monophones occur in the numbers part of the corpus. The remaining 7 nodes were not used. Each monophone model uses one HMM state, which is repeated three to six times, depending on the respective monophone. For the triphone-based alignment we substituted in the monophone-based alignment each monophone label by a new label which depended on both the monophone's left and right context. This gave us a set of 151 triphone labels (word-internal only). This alignment was then used to train the context-dependent MLPs which have 151 output nodes. The triphone HMM models use 3 states for duration modeling. Only the silence model uses just one state without duration modeling.

The results of the three one-stream systems are given in Table 2. In order to evaluate whether a difference in Word Error Rate (WER) is significant, we carried out a significance test at a confidence level of 97.5%. A result is therefore significantly different from the best result achieved (that is, for CI models: 7.2; for CD models: 6.6) if it lies outside the interval of [6.61,7.79] for the CI models and [6.04,7.16] for the CD models.

	CI models	CD models
RASTA	8.0	8.4
PLP	7.2	6.6
MSG	7.3	6.8

Table 2. WERs of each of the three feature streams as employed in a standard (one-stream) recognizer.

3.1. Feature Concatenation

We first investigate MS feature concatenation, for which each feature stream was concatenated to each other feature stream, and an acoustic model was trained on the combined stream. This leads to an increased input layer size, but the hidden and output layer sizes can be kept the same. Concatenation of the 3 streams leads to a rather large feature stream which needs to be processed. This might lead to a problem in standard Gaussian modeling, increasing the number of the necessary Gaussians and producing a large number of HMM parameters. In a hybrid system, we can afford to have a large input feature vector as the modeling is carried *between* the input and the output layer of the MLP, so that the latter, which is responsible for the number of mixture weights per HMM state, does not change in size.

	CI models	CD models
RASTA-PLP	5.5	6.3
PLP-MSG	5.1	5.5
RASTA-MSG	5.1	6.1
RASTA-PLP-MSG	4.7	6.2

Table 3. WERs for the MS systems employing feature combination.

The results are given in Table 3. Although the RASTA feature stream when used by itself is significantly worse than the other two streams (cf. Table 2), after feature combination each concatenated

feature stream leads to a significantly improved recognition rate. In the case of the CI models, best results were achieved when all three streams were concatenated. For the CD models, it was the combined PLP-MSG stream which gave the best results.

3.2. Probability Combination

In MS probability combination we investigated combinations according to Equations (1) (product rule), (3) (standard sum rule) and (5) (FC SUM rule). We use equal weights (for each class and expert) in all experiments. For FC processing, also the MLPs from feature combination were employed as defined by Equation (5).

	CI models	CD models
RASTA∗PLP PRODUCT	7.3	6.8
RASTA+PLP SUM	6.1	5.8
RASTA-PLP FC SUM	5.2	5.7
PLP∗MSG PRODUCT	7.8	6.7
PLP+MSG SUM	5.9	5.6
PLP-MSG FC SUM	5.0	5.6
RASTA∗MSG PRODUCT	7.5	6.5
RASTA+MSG SUM	6.2	6.4
RASTA-MSG FC SUM	5.1	5.9
RASTA∗PLP∗MSG PROD.	7.5	8.2
RASTA+PLP+MSG SUM	5.7	5.7
RASTA-PLP-MSG FC SUM	4.5	5.7

Table 4. WERs of the MS systems employing probability combination with two and three different feature streams.

In Table 4, we can see the clear tendency that the (standard) sum rule always outperforms the product rule (for both the CI and the CD models). The sum rule achieved significantly improved results as compared to the single stream results, though it could not improve over the respective feature concatenation when using the context-independent models. For the context-dependent case, the sum rule outperformed feature combination in half of the cases (i.e. for RASTA+PLP and RASTA+PLP+MSG).

With FC processing we are able to further enhance performance achieving the best results for each respective combination of streams, when employing CI models. With CD modeling there are two cases (PLP-MSG and RASTA-PLP-MSG) where the FC SUM is not better than the standard sum or even feature combination due to the good results of the combined PLP-MSG feature stream.

Significantly lowest WER (4.5) was achieved with FC MS processing employing all three feature streams and CI modeling.

4. CONCLUSION

With the different characteristics of the PLP and the MSG feature streams, these streams are especially well suited for MS processing. The RASTA features differ from the PLP features only in the additional RASTA-filter but this difference is still strong enough (even on our telephone-recorded but otherwise clean utterances) to be exploited in MS processing to improve recognition performance. Feature concatenation of all three streams led to good results which could be further improved when using FC probability combination. With the FC SUM we achieved the best results for each combination of features.

Context-dependent modeling using word-internal triphones improved results mainly of the one-stream systems and in standard probability combination, due to better modeling capabilities of the context-dependent HMMs. Feature concatenation in these systems resulted in smaller improvements. For this reason it was harder to achieve a significant gain from MS FC processing. Cross-word triphones might be needed additionally to the word-internal triphones to enhance improvement.

A disadvantage of using CD models can be reduced generalization ability and lack of robustness due to sparsity in training data. To circumvent this problem, HMM based system train models at many different levels of context, such as monophone, biphone and triphone models [7], which are then linearly smoothed. Such an approach is not possible in HMM/MLP hybrid systems. Instead, we plan on using the MS approach as a way to "back-off" our context-dependent hybrid system with a context-independent system.

ACKNOWLEDGMENTS

Astrid Hagen was supported by the Portuguese FCT (Fundação para a Ciência e a Tecnologia) scholarship SFRH/BPD/6757/2001. Additionally, this work was partially funded by the FCT project POSI/33846/PLP/2000. INESC-ID Lisbon had support from the POSI Program of "Quadro Comunitário de Apoio III".

5. REFERENCES

[1] S. Okawa, E. Bocchieri, and A. Potamianos, "Multi-band speech recognition in noisy environment," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 2, pp. 641–644, 1998.

[2] Dan P.W. Ellis, "Stream combination before and/or after the acoustic model," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 3, pp. 1635–1638, 2000.

[3] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination to noise robust ASR," *Speech Communication*, vol. 34, no. 1-2, pp. 25–40, 2001.

[4] K. Kirchhoff, G.A. Fink, and G. Sagerer, "Conversational speech recognition using acoustic and articulatory input," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 3, pp. 1435–1438, 2000.

[5] H. Meinedo and J.P. Neto, "Combination of acoustic models in continuous speech recognition hybrid systems," *Int. Conf. on Spoken Language Processing*, vol. 2, pp. 931–934, 2000.

[6] Astrid Hagen, *Robust speech recognition based on multi-stream processing*, Ph.D. thesis, Département d'informatique, École Polytechnique Fédérale de Lausanne, Switzerland, 2001.

[7] K. Lee, "Context-dependent phonetic hidden markov models for speaker-independence continuous speech recognition," in *IEEE Trans. on Acoustics, Speech and Signal Processing*, 1990, vol. 38, pp. 599–609.