FRAME-BASED ACOUSTIC FEATURE INTEGRATION FOR SPEECH UNDERSTANDING

Loic Barrault, Christophe Servan, Driss Matrouf, Georges Linarès, Renato De Mori

LIA
University of Avignon, BP 1228
84911 Avignon Cedex 9 - France
{loic.barrault, christophe.servan, driss.matrouf, georges.linares, renato.demori}@univ-avignon.fr

ABSTRACT

With the purpose of improving Spoken Language Understanding (SLU) performance, a combination of different acoustic speech recognition (ASR) systems is proposed. State *a-posteriori* probabilities obtained with systems using different acoustic feature sets are combined with log-linear interpolation. In order to perform a coherent combination of these probabilities, acoustic models must have the same topology (*i.e.* same set of states). For this purpose, a fast and efficient *twin* model training protocol is proposed. By a wise choice of acoustic feature sets and log-linear interpolation of their likelihood ratios, a substantial Concept Error Rate (CER) reduction has been observed on the test part of the French MEDIA corpus.

Index Terms— speech recognition, posterior probabilities combination, speech understanding, frame based combination

1. INTRODUCTION

It is known that Automatic Speech Recognition (ASR) systems make errors that limit the potential for their application.This is due to the imperfection of the models used, to limitations of the features extracted and the approximations performed by the recognition engines. With the purpose of increasing robustness, it has been proposed to combine the results of different ASR systems. Attempts have been reported [1] on the use of neural networks, decision trees and other machine learning techniques to combine the results of ASR systems, or components of them, fed by different feature streams or using different models in order to reduce WER. In [2] it is shown that log-linear combination provides good results when used for integrating probabilities provided by acoustic models.

Complementary system combination and related problems are reviewed and discussed in [3]. It is noticed that different systems may lead to performance improvements especially if systems are truly complementary.

The use of different features for different classes has been suggested in [4]. This is motivated by the assumption that some characteristics that are de-emphasized by a particular feature are emphasized by another feature, and therefore the combined feature streams capture complementary information present in individual features. Along the same line, some specific parameters have been integrated into a single stream of features [5]. A generalization of this approach consists in concatenating different sets of acoustic features into a single stream. Another approach, consisting in frame based system combination is proposed in [6]. It is shown that the corresponding decoding process compares favorably to decoding based on confusion network combination.

An aspect which has not been investigated yet is the improvement of the performance of a Spoken Language understanding (SLU) system by using different acoustic feature sets for conceptual decoding. The process uses conceptual language models to extract meaning from a lattice of word hypotheses ([7]). If the feature sets are sufficiently different, it is more likely that semantic important words are hypothesized in a word lattice and are used by a meaning extraction method that makes decisions based on conceptual consistency and not on just word accuracies. Experimental evidence is provided that the choice of feature sets as well as the combination methods proposed here provides consistent recognition and interpretation improvements with respect to the use of a single feature stream.

Frame-based posterior probabilities combination is computed before decoding using multi-stream framework as proposed, for example, in [8]. In order to combine posterior probabilities, sub-systems are considered which have equal topology (*i.e.* same set of states). A training technique ensuring model consistency is used to allow coherent probability combination without pseudo-states for synchronism. Rather than using first and second time derivatives as different streams, in the proposed system, three fairly different speech analysis methods are used. The features used are Perceptual Linear Prediction (PLP) coefficients [9], PLP with JRASTA filtering [10] and Multi Resolution Analysis computed as described in [11]. Each stream includes first and second time derivatives. This is motivated by the fact that these different speech analysis methods provide different recognition perfor-

mance in different zones of the acoustic space.

In the n^{th} speech frame, a feature vector Y_n^i is computed for the i^{th} feature set and its derivatives. A state likelihood $L(Y_n^i|q)$ is then computed for each state q. Likelihoods are normalized and combined frame-by-frame to produce a composed normalized likelihood ratio. Given a set of feature sets, many sub-systems can be built and their results can be combined in various ways. Log-linear interpolation is performed on likelihood ratios as suggested in [12].

Section 2 describes sub-system architectures and the specific training procedure used for combining the estimation of their parameters. Log-linear combination of likelihood ratios is presented in Section 3. Section 4 reports experimental results.

2. SYSTEM ARCHITECTURE AND *"TWIN"* MODEL TRAINING

Speech generation is a source of information producing a signal in which symbols are encoded. Given a sampled input signal $S = \{s(k\tau)\}$, where τ is the sampling period, let us consider the sequence of samples in a time window of length T and represent such a sequence for the n^{th} window as: $Y_n = [s(k\tau)]_{nT}^{(n+1)T}, n = 0, \ldots, N$. Feature vectors are used for computing likelihoods about the presence in a signal frame of symbols q of a vocabulary Q. Let us consider $\Im^i, i = \{1, \ldots, I\}$, a set of acoustic spaces which are realizations of the acoustic space $\Im$ corresponding to different feature sets $\{Y^i\}$, and Y_n^i the instances of the frame Y_n in those acoustic spaces. Let us consider context-dependent acoustic models made of Hidden Markov Models (HMM) in which a gaussian mixture (GMM) models the probability density for each state represented by a symbol q.

Generation of speech hypotheses is performed by a decoding strategy which evaluates sequences of model states using probabilities about states and frame features; an example of which is the state posterior probability $P(q|Y_n)$. In order to combine multiple feature sets, an efficient training technique which preserves the topology of the acoustic models is needed. For this purpose, instead of training acoustic models separately, a twin model training strategy which preserves the topology of the acoustic models is proposed. Let us consider a source model M^0 trained with feature set Y^0.

Our goal is to create new *twin* models M^i, which uses acoustic feature set Y^i, having the same set of states as M^0. To do so, forced alignment of trainning corpus with M^0 is used. Each GMM associated with each state in M^i, is trained using the following steps:

- Expectation step of the EM algorithm is performed using feature set Y^0 on corresponding GMM of M^0.
- The Maximization step of the EM algorithm is performed using feature Y^i with model M^i.
- Re-estimate M^i using some iterations of maximum *a-posteriori* (MAP) adaptation. The segmentation of training corpus is updated using the model M^i obtained at each iteration.

3. FRAME BASED FEATURE COMBINATION OF POSTERIOR PROBABILITIES

After different acoustic models have been trained, they are used in the architecture represented in Figure 1. Likelihoods $L(Y_n^i|q)$ are computed synchronously for each feature set. Then, for each frame, an integrated likelihood ratio $LR(Y_n, q)$ is computed using different possible combinations and integrations as discussed in the next section. Several ways

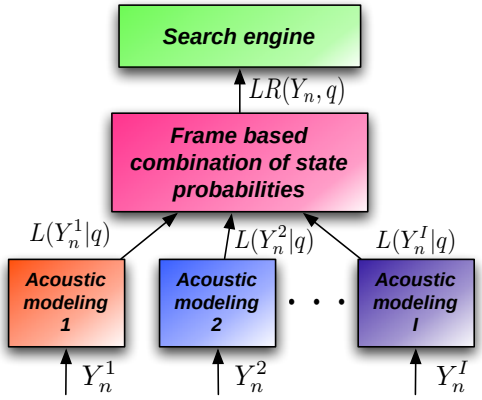

Fig. 1. *Architecture for frame-based feature combination.*

for combining posterior probabilities can be considered. For combining frame-based probabilities, it is assumed that the system is in state q if all feature sets agree on this. Assuming statistical independence among feature sets, one gets:

$$\hat{P}(q|Y_n) = P(q|Y_n^1, \ldots, Y_n^I) = \prod_{i=1}^{I} P(q|Y_n^i) \qquad (1)$$

The following log-linear combination of likelihood ratios is used:

$$LLCLR(n, q) = \sum_{i=1}^{I} \alpha_i \log \left[\frac{L(Y_n^i|q)}{\sum_{g \in Q} L(Y_n^i|g)} \right] \qquad (2)$$

If there is no reason for distinguishing among acoustic feature sets, the following assumption can be made: $\alpha_i = \frac{1}{I}$. This assumption was found to produce good results in the experiments described in the next section.

4. EXPERIMENTS

The system used is SPEERAL, the HMM based ASR system developed at LIA. It has 64 Kword vocabulary, 10040 cross-word context-dependent models, 3600 emitting states tied using decision-tree method and 232716 gaussian components. The acoustic models of the systems were trained separately, using the twin model training approach, for each feature sets using 82 hours of telephone speech of the French corpus ESTER The train set, with 82639 words, of another French corpus MEDIA was used for adaptation.Three feature sets were considered corresponding to fairly different ways of transforming the speech sample sequences. The first feature set is a vector of Perceptual Linear Prediction (PLP) coefficients. The second feature set is obtained by adding RASTA filtering to the PLP features (RPLP). The third feature set is computed with Multi Resolution Analysis (MRA) followed by Principal Component Analysis. All feature vectors also contains first and second time derivatives. A set of results in terms of Word Error Rate (WER), reported in Table 1, were obtained with the test part of the MEDIA corpus. MEDIA is a 1250 dialogue corpus recorded using the Wizard of Oz protocol : 250 speakers made hotel reservations following 5 different scenarios. This corpus of telephone speech consists of 3769 sentences and 25482 words. It has been manually transcribed and conceptually annotated according to the semantic representation defined within the project and presented in [13]. The test part of MEDIA corpus is composed of 83 different concept labels for a total of 8373 concepts in 200 dialogues. Results with the frame-based combination (LLC) of MRA, RPLP and PLP are reported. The confidence interval is also reported for each result. A WER reduction of more than

Table 1. *Percentage results on the MEDIA test corpus (3769 sentences and 25482 words).*

Feature set	WER (%)	Conf. interval (%)
MRA	33.9	0.58
RPLP	32.8	0.58
PLP	32.8	0.58
LLC	28.1	0.55

14% relative to the best system using only one feature set was observed. Table 2 reports the Concept Error Rates (CER) obtained with each feature set and their log-linear combination. A lattice of concept hypotheses is generated from a lattice of word hypotheses as described in [7].

Table 2. *Concept Error Rates obtained with the 1-best concept hypothesis (%).*

	MRA	RPLP	PLP	LLC
CER (%)	37.0	37.1	35.1	32.4

The results of Table 2 refer to the 1-best concept sequence. Oracle results are obtained by extracting the N-best list of concept hypotheses from the concept lattice. The oracle concept error rate is reported in Figure 2 as function of N. Figure

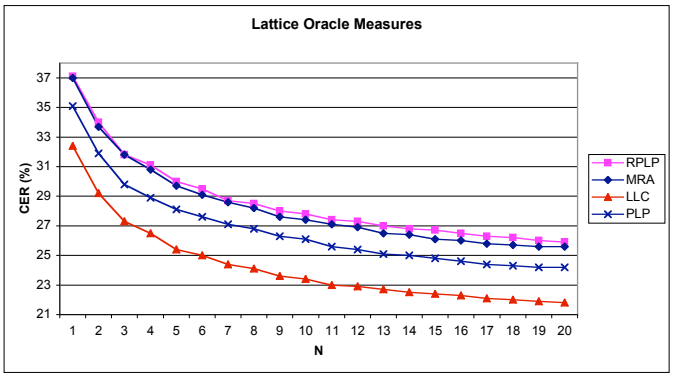

Fig. 2. *Oracle CER.*

2 shows the same general trend for the four systems. The performance improvement is relatively stable for all values of N and varies between 7% to 10% relatively to the best system using a single feature set (PLP).

Table 3 reports the CER when there is a consensus between the concept hypotheses generated by the best feature set (PLP) and the combination (column #3). Consensus appears in more than 71% of the turns and in this case, a strong reduction of CER is observed. Consensus appears to be a valid confidence indicator.

Table 3. *Relation between conceptual recognition performance and transcription performance (2992 turns in total).*

	LLC Best	Same hyp.	LLC Worst	Total
Comparison bewteen LLC and PLP				
% turns	13.1	71.5	8.7	100
CER LLC	35.5	**24.7**	57.8	32.4
CER PLP	61.6	**24.7**	30.9	35.1
WER LLC	32.6	**22.7**	36.1	28.1
WER PLP	42.8	**26.1**	35.2	32.8
Comparison bewteen LLC and RPLP				
% turns	15.1	70.7	7.4	100
CER LLC	30,9	**27,3**	56,4	32,4
CER RPLP	60,2	**27,3**	34	37,1
WER LLC	30,7	**23,2**	36,5	28,1
WER RPLP	41,7	**25,8**	36	32,8
Comparison bewteen LLC and MRA				
% turns	15.9	69.5	7.7	100
CER LLC	32,1	**26,7**	56,8	32,4
CER MRA	60	**26,7**	32,1	37.0
WER LLC	30,4	**23,8**	35,3	28,1
WER MRA	39,6	**28.0**	38,9	33,9

Results obtain for SLU suggests a few comments. Evidence is shown that using multiple feature streams provide

substantial CER reduction. Comparison of feature sets with the combination shows that when different systems give the same conceptual hypotheses, then it is likely that they are correct. A low WER is obtained on sentences where both models provide the same interpretation. This demonstrates that consensus among concept hypotheses obtained with multiple systems is a good confidence indicator for both speech recognition and speech understanding.

A different behavior can be observed for sentences where one feature performs better than the combination. For sentences where a single feature set gives better conceptual recognition hypotheses than LLC, a recognition rate far better than in the other cases is observed. For PLP and RPLP, WER for sentences where they provide better conceptual recognition results is even **lower** than the one obtained with LLC. A detailed analysis of these sentences should give a lot of information about what acoustic events cause both speech recognition and understanding errors.

5. CONCLUSION

The improvement of speech understanding using complementary systems is presented in this paper. The use of different feature sets, based on different speech analysis methods provides good speech understanding performance. Experimental results show that frame based combination leads to substantial error reduction in speech understanding. In particular, a CER reduction of more than 14.3% has been observed on the test part of MEDIA corpus. This shows that a wise choice of acoustic feature sets to be combined has a positive impact in Speech Understanding results.

As a perspective, the use of word lattice at the input of the conceptual decoder instead of just the 1-best should generate new conceptual hypotheses and lead to better performance. It is likely that the 1-best transcription hypothesis not always gives the best conceptual hypothesis that could be obtained with the whole word lattice.

6. REFERENCES

[1] B. Zhang, S. Matsoukas, and R. Schwartz, "Discriminatively trained region dependent feature transforms for speech recognition," in *IEEE International Conference on Acoustics, Speech and Language Processing*, Toulouse, France, 2006.

[2] A. Zolnay, R. Schluter, and H. Ney, "Acoustic feature combination for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Language Processing*, Philadelphia, PA, March 2005, vol. 1, pp. 457–460.

[3] M. J. F. Gales, D. Y. Kim, P. C. Woodland, H. Y. Chan, D. Mrva, R. Sinha, and S. E. Trante, "Progress in the cu-htk broadcast news transcription system," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14(5), pp. 1513–1525, september 2006.

[4] A. K. Halberstadt and J. R. Glass, "Heterogeneous measurements and multiple classifiers for speech recognition," in *International Conference on Spoken Language Processing, Interspeech*, Sydney, Australia, 1998, p. 13791382.

[5] A. Zolnay, R. Schluter, and H. Ney, "Robust speech recognition using a voiced-unvoiced feature," in *International Conference on Spoken Language Processing, Interspeech*, Denver, CO, 2002, vol. 2, pp. 1065–1068.

[6] B. Hoffmeister, T. Klein, R. Schluter, and H. Ney, "Frame based system combination and a comparison with weighted rover and cnc," in *International Conference on Spoken Language Processing, Interspeech*, 2006, pp. 537–540.

[7] Christian Raymond, Frédéric Béchet, Renato De Mori, and Géraldine Damnati, "On the use of finite state transducers for semantic interpretation," *Speech Communication*, vol. 48, no. 3-4, pp. 288–304, 2006.

[8] H. Bourlard and S. Dupont, "Sub-band based speech recognition," in *IEEE International Conference on Acoustics, Speech and Language Processing*, Munich, Germany, 1997, pp. 1251–1254.

[9] Hynek Hermansky, "Perceptual linear predictive (plp) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, 1990.

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

[11] R. Gemello, F. Mana, D. Albesano, and R. De Mori, "Multiple resolution analysis for robust automatic speech recognition," *Computer Speech and Language*, vol. 20, no. 1, pp. 2–21, 2006.

[12] O. Siohan, B. Ramabhadran, and B. Kingsbury, "Constructing ensembles of asr systems using randomized decision trees," in *IEEE International Conference on Acoustics, Speech and Language Processing*, Philadelphia, PA, March 2005, vol. 1, pp. 197–200.

[13] Christophe Servan, Christian Raymond, Frédéric Béchet, and Pascal Nocéra, "Conceptual decoding from word lattices: application to the spoken coprus media," in *International Conference on Spoken Language Processing, Interspeech*, September 2006, p. 4.