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 Word Error Rates (WER). In [2] it is shown that log-linear combination provides good results when used for integrating probabilities computed with acoustic models.

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 acoustic features to characterize different phoneme classes has been suggested in [4]. Along this line, some specific parameters or different sets of acoustic features have been integrated into a single stream of features [5]. 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 possibility of improving the performance of a Spoken Language Understanding (SLU) system by using different acoustic feature sets for conceptual decoding. This process uses conceptual language models to extract meaning from a lattice of word hypotheses [7]. If the feature sets contribute with different hypotheses to phoneme equivocation, it is likely that a suitable combination of them leads to an equivocation reduction. As a consequence of this, it is more likely that semantically 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 result in consistent recognition and interpretation improvements with respect to the use of a single feature stream.

Frame-based probability combination is computed before decoding as proposed, for example, in multi-stream framework [8]. In order to combine posterior probabilities, subsystem models 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 introducing pseudo-states for monitoring synchronism. Rather than using first and second time derivatives as different streams, in the proposed system, three fairly different feature sets are used. They are Perceptual Linear Prediction (PLP) coefficients [9], PLP with JRASTA filtering [10] and Multi Resolution Analysis (MRA) computed as described in [11]. Each stream includes first and second time derivatives.

This work is supported by the 6th Framework Research Programme of the European Union (EU), Project LUNA, IST contract no 33549. For more information about the LUNA project, please visit www.ist-luna.eu.

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. Log-linear interpolation is performed as suggested in [2] on likelihood ratios.

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 corresponding to different feature sets $\{Y^i\}$, and $Y_n^i, i = \{1, \ldots, I\}$ be the instances of the frame Y_n in each acoustic space. Let us consider context-dependent acoustic models made of Hidden Markov Models (HMM) in which a gaussian mixture models the probability density for each state represented by a symbol q.

Generation of word hypotheses is performed by a decoding strategy which approximates posterior probabilities $P(q|Y_n)$ of model states by likelihood ratios.In order to combine multiple feature sets, an efficient training technique which preserves the topology of the acoustic models is proposed. Instead of training acoustic models separately, a twin model training strategy is used. Let us consider a source model M^0 trained with feature set Y^0. The goal is to create new *twin* models M^i with the same set of states as M^0, which use acoustic feature sets Y^i. To do so, forced alignment of the trainning corpus is performed with M^0. Each gaussian mixture model (GMM) associated with each state in M^i, is trained using the EM algorithm with the following steps:

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

3. FRAME BASED FEATURE COMBINATION OF POSTERIOR PROBABILITIES

The models 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 $LR(n, q)$ is computed. Several ways for combining posterior probabilities can be considered. The following computation based on log-linear combination of likelihood ratios has been found to produce good results in the experiments described in the next section.

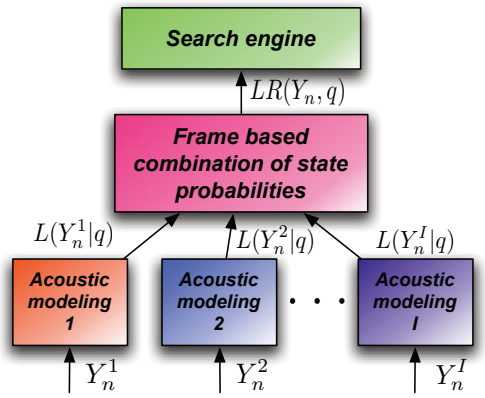

Fig. 1. *Architecture for frame-based feature combination.*

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

Attempts to perform linear combination and to estimate fixed or varying values of α_i did not produce significant improvements with respect to the simple choice of using the (1) with $\alpha_i = \frac{1}{I}$. This choice corresponds to just multiply probabilities obtained with different feature sets. This is expected to reduce phoneme equivocation if probabilities of different sets tend to be comparable for states of the phoneme model that has been uttered and very different for states of the other phonemes. In the latter case, the resulting probability would be strongly reduced with respect to the highest among the probabilities obtained with different sets. Such a situation is expected to reduce phoneme equivocation in a way that primarily depends on the choice of feature sets.

Instead of combining likelihood ratios, it is possible to concatenate different sets of acoustic features into a single stream. In order to reduce modeling complexity and problems due to data sparseness, algorithms have been described to select subsets of features in a long stream using a criterion that optimizes automatic classification of speech data into

phonemes or phonetic features. Unfortunately, pertinent algorithms are computationally intractable with these types of classes as stated in [12], where a sub-optimal solution is proposed. Such a solution ignore some features by selecting a set of acoustic measurement that guarantees a high value of the mutual information between acoustic measurements and phonetic distinctive features.

4. EXPERIMENTS

The HMM based ASR system used for the experiments described in this section is SPEERAL [13]. It has 64 Kword vocabulary, 10040 cross-word context-dependent models, 3600 emitting states tied using a decision-tree method and 232716 gaussian components. The acoustic models were trained separately 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, namely PLP, RASTA-PLP and MRA followed by Principal Component Analysis. All feature vectors also contains first and second time derivatives. A set of results obtained with log-linear combination (LLC) are reported in Table 1. They 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 a semantic representation [13]. The test part of the MEDIA corpus is composed of 83 different concept labels for a total of 8373 concepts in 200 dialogues.

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

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

A WER reduction of more than 14% relative to the best system using only one feature set was observed.

Table 2 shows the Concept Error Rates (CER) obtained with each feature set and their log-linear combination.

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

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

A lattice of concept hypotheses is generated from a lattice of word hypotheses as described in [7]. 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 and selecting the one that minimize the CER among these hypotheses. The oracle concept error rate is reported in Figure 2 as function of N. Figure 2 shows

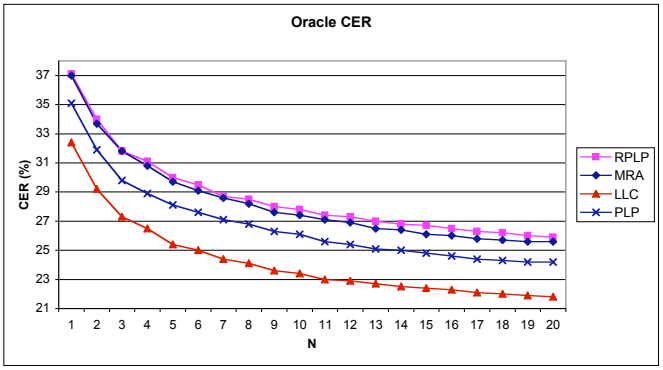

Fig. 2. *Evolution of oracle CER as function of the depth N of the N-best list of concept hypotheses.*

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 values of CER and WER for different situations. Column *"LLC Better"* corresponds to the situations in which LLC has provided, in a dialog turn, a lower CER than one of the features mentioned in the line below. The column *"LLC Worse"* corresponds to the opposite. The column *"Consensus"* corresponds to a consensus situation between LLC and the feature in the next line. It appears that there is consensus between LLC and PLP in more than 71% of the dialog turns. In this case a strong CER reduction is observed. Thus, consensus appears to be a valid confidence indicator.

Evidence is shown that using multiple streams of suitable features provide substantial CER reduction. When different feature sets lead to the same conceptual hypotheses it is likely that the hypotheses are correct. A low WER is obtained on sentences where feature sets provide the same interpretation. Consensus among concept hypotheses obtained with multiple features is a good confidence indicator for speech recognition as well as for speech understanding.

A different behavior is 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, for sentences where a single feature set provides better conceptual recognition results, WER is even **lower** than the one obtained with LLC. A detailed analysis of these cases make evident acoustic events which cause both speech recognition and understanding errors.

	LLC Better	Consensus	LLC Worse	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

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

5. CONCLUSION

Significant performance improvements in speech understanding has been obtained using frame based linear combination of well chosen acoustic feature sets. 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. It has also been observed that most of the errors appear when there is no consensus between conceptual hypotheses generated with different feature sets.

As a perspective, the use of a word lattice at the input of the conceptual decoder instead of just the 1-best hypothesis should increase the probability of obtaining the correct semantic interpretation by allowing structural constraints.

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, pp. I–I.

[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, pp. 1379–1382.

[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] C. Raymond, F. Béchet, R. De Mori, and G. 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] H. 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] M. Kamal Omar and M. Hasegawa-Jonhson, "Maximum mutual information based acoustic-features representation of phonological features for speech recognition," in *IEEE International Conference on Acoustics, Speech and Language Processing*, Orlando, FL, 2002, vol. 1, pp. 81–84.

[13] C. Servan, C. Raymond, F. Béchet, and P. Nocéra, "Conceptual decoding from word lattices: application to the spoken corpus media," in *International Conference on Spoken Language Processing, Interspeech*, September 2006, pp. 1614–1617.