DYNAMIC SELECTION OF ACOUSTIC FEATURES IN AN AUTOMATIC SPEECH RECOGNITION SYSTEM

Loic Barrault[1], *Driss Matrouf*[1], *Renato De Mori*[1], *Roberto Gemello*[2] *and Franco Mana*[2]

[1]LIA, BP 1228
84911 Avignon Cedex 9 - France
loic.barrault,renato.demori,driss.matrouf@univ-avignon.fr

[2]LOQUENDO
Via Valdellatorre, 4 - 10149 Torino - Italy
roberto.gemello,franco.mana@loquendo.com

ABSTRACT

A general approach for integrating different acoustic feature sets and acoustic models is presented. A strategy for using a feature set as a reference and for scheduling the execution of other feature sets is introduced. The strategy is based on the introduction of feature variability states. Each phoneme of a word hypothesis is assigned one of such states. The probability that a word hypothesis is incorrect given the sequence of its variability states is computed and used for deciding the introduction of new features.

Significant WER reductions have been observed on the test sets of the AURORA3 corpus. Using the CH1 portions of the test sets of the Italian and Spanish corpora, word error rate reductions respectively of 16.42% for the Italian and 29.4% for Spanish were observed.

1. INTRODUCTION

It is known that Automatic Speech Recognition (ASR) systems make errors (see, for example, [8]). This is due to the imperfection of the various models used, on the limitations of the feature extracted and on the approximations of the recognition engines.

With the purpose of increasing robustness, recent ASR systems combine streams of different acoustic measurements, such as multi-resolution spectral/time correlates ([2], [12], [7], [4]).

Other approaches integrate some specific parameters into a single stream of features ([11], [13]). A generalization of this approach consists in concatenating different sets of acoustic features into a single stream. In order to reduce modelling complexity, 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 [5], where a sub-optimal solution is proposed. Such a solution consists in selecting a set of acoustic measurement that guarantees a high value of the mutual information between acoustic measurements and phonetic distinctive features.

The approach described in this paper consieders the possibility of dynamically combining different feature sets and acoustic models in an ASR system.
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 follows:

$$Y_n \quad = \quad [s(k\tau)]_{nT}^{(n+1)T}, n = 0, 1, ..., N$$

For each value of n, the window sequence $[s(k\tau)]_{nT}^{(n+1)T}$ of signal samples is transformed into a feature vector $Y^a(nT)$ represented in a feature space $\Im^a$.
Features have an intrinsic variability with respect to the symbols $q \in Q$ which describe a spoken message. Variability may cause equivocation represented by the fact that different symbols of Q may be coded into signal segments leading to the same vector $Y^a(nT)$. Equivocation varies from point to point of a given feature space. In order to reduce equivocation, it is thus interesting to vary the choice or the use of features depending on the sample sequence based on which symbol hypotheses are hypothesized.

2. USING DIFFERENT FEATURES AND MODELS

Given feature samples $Y^a(nT)$, hypotheses about symbols $q \in Q$ are generated by computing the posterior probabilities $P_\mu[q|Y^a(nT)]$. Symbols may represent phonemes, phonemes in context, transients or other phonetic descriptors. Computation of these probabilities is performed using acoustic models μ. If different models and features are available, then posterior probabilities can be obtained with log-linear interpolation as follows:

$$\log P[q|Y_n] = \sum_{\mu,a} w_\mu^a[Y^r(nT)] \log P_\mu[q|Y^a(nT)] \qquad (1)$$

where $w_\mu^a[Y^r(nT)]$ are weights depending on the feature sample in a reference space indicated by the super-script r. Initially, speech analysis is performed in the reference space $\Im^r$. The reference features can be the ones that produce the best ASR results or good results with minimal computation time.

Depending on the value of $Y^r(nT)$, other feature sets can be used for performing a more accurate analysis and for reducing the equivocation on the generation of hypotheses about symbols $q \in Q$.

Equivocation between a channel source S which emits symbols $f \in Q$ and the receiver R which hypothesizes symbols $qg \in Q$ is defined as follows ([9]):

$$H_R(S) = -\sum_{f,g} P[f, g] \log P[g|f] \qquad (2)$$

The (2) is an overall definition of equivocation over the entire acoustic space of one or more feature sets.

Given a specific window sequence of signal samples Y_n and one or more sets of acoustic features, if f is the most likely phoneme hypothesis for the specific time frame, then the probability of equivocation for that frame is given by:

$$P_{eq}(nT) \quad = \quad \sum_{g \neq f} P[g|Y_n]$$

As an example of the application of the just introduced concepts, a short introduction of the system used for performing the experiments described in this paper is provided in the following.

Two versions of an hybrid system consisting of an Artificial Neural Network (ANN) and a set of Hidden Markov acoustic Models (HMM) are used. The two recognizers are fed by different sets of acoustic features, but have the same topology.
The feature sets are those obtained by Multi Resolution Analysis (MRA) followed by Principal Component Analysis (PCA) and by Perceptual Linear Prediction (PLP) followed by RASTA filtering. The latter features will be indicated as JRASTAPLP. The same denoising technique is used for both feature sets (Gemello). The ANNs are trained to recognize phonemes and transitions using a corpus of phonetically balanced sentences which are completely independent from the test data.

It is important to notice that combining the scores of the two recognizers does not change the recognition results if their most likely hypotheses are the same. The ANNs have 636 outputs, one for each phoneme and each transition between two successive phonemes. Two streams of acoustic feature, namely $Y^m(nT)$ and $Y^j(nT)$ are generated. Indices m and j respectively refer to MRA features and JRASTAPLP features. The vectors $Y^m(nT)$ and $Y^j(nT)$ represent two different observations of a speech segment Y_n centered on the same sample. From now on, these two indices will be used to indicate the two types of features. The value of T is 10 msecs and each feature set contains seven analysis frames centered on the frame at nT.

The two recognizers have acoustic models which induce probability distributions in the acoustic spaces $\Im^m$ and $\Im^j$ of the two feature types. For each space, the distributions are posterior probabilities of a phoneme f or a diphone representing the transition between two phonemes given an observation in that space. The *variability* of the features extracted can be described using the osterior probabilities for phonemes and diphones in a point of an acoustic space. Vectors $Y^m(nT)$ and $Y^j(nT)$ may have similar or very different posterior probabilities for the same utterance of phoneme f.

Let us consider, for the sake of simplicity, only phoneme symbols indicated by q. The 1 is rewritten making explicit reference to the features obtained with multi-resolution analysis, indicated as $Y^m(nT)$ and two acoustic models, one based on an ANN, indicated by A, and one based on Gaussian mixtures, indicated by G. The 1 can thus be written as follows:

$$\begin{aligned} \log P[q|Y_n] \quad &= \quad w_A^m[Y^m(nT)] \log P_A[q|Y^m(nT)] \qquad (3) \\ &+ \quad w_G^m[Y^m(nT)] \log P_G[q|Y^m(nT)] + \Re^m(nT) \end{aligned}$$

$\Re^m(nT)$ is a term that represents the introduction of additional features which can reduce the equivocation in the point of the acoustic space $\Im^r = \Im^m$ corresponding to the n-th time frame. Different feature spaces or even different granularities in the same feature space can be used in $\Re^m(nT)$. The weights in the 4 depend on a specific time frame and can also be set to binary values allowing the system to switche between feature sets or acoustic models.

The contribution of $\Re^m(nT)$ can be neglected if, based on the available feature sets and models, this will have little effect on the equivocation of the phoneme hypothesization process. In other words, if the variability of feature $Y^m(nT)$ is low and the feature corresponds almost always to the same phoneme.

In a previous paper [1], it is proposed to assign a reliability state σ to each vector $Y^m(nT)$ in the reference space. The assignment is now described.

A comparison between phoneme posterior probability distributions obtained with ANN and GMM is performed on a segment $SEG_a(b, e, t)$. Symbol a describes the type of features, b indicates the beginning time of the segment, e indicates the end time and t the time at the middle. The discrepancy between the two posterior probability distributions $P_A^a[q|SEG_a(b, e, t)]$ and $P_G^a[q|SEG_a(b, e, t)]$ for $q \in Q$, is the Kullback-Leibler distance (KLD) indicated as:

$$\begin{aligned} KLD_a[SEG_a(b, e, t)] \quad &= \\ = D\left[P_A^a[q|SEG_a(b, e, t)] || P_G^a[q|SEG_a(b, e, t)]\right] \quad &= \\ = \sum_{g \in Q} P_A^a[g|SEG_a(b, e, t)] \log \frac{P_A^a[g|SEG_a(b, e, t)]}{P_G^a[g|SEG_a(b, e, t)]} \qquad (4) \end{aligned}$$

The symbol with the highest posterior probability is considered as the hypothesis generated with each model in the given segment. These hypotheses are respectively indicated as $g_A^a(b, e, t)$ and $g_G^a(b, e, t)$. In principle, a low value of $KLD_a[SEG_a(b, e, t)]$ does not necessarily indicate that the two probability distributions propose the same value of q with the highest posterior probability. For this reason, an indicator of good modeling and reliable decision is the truth of the predicate:

$$\begin{aligned} CONS_a[SEG_a(b, e, t)] = KLD_a[SEG_a(b, e, t)] < 0.4 \wedge \\ \wedge \{g_A[SEG_a(b, e, t)] = g_G[SEG_a(b, e, t)]\} \end{aligned}$$

for which there is a high probability that the phoneme with the highest probability is correct. This is the condition in which the first contribution to equivocation is computed.

Let us define a variability/reliability state σ_1 corresponding to the case $CONS_a[SEG_a(b, e, t)] = true$. Given a corpus, characterizing an application domain, if it is observed that equivocation is low in σ_1 when $a = m$, then the term $\Re^m(nT)$ can be ignored in this state. When $CONS_a[SEG_a(b, e, t)]$ is not true, either better models or other features should be considered. Assuming that the available models are already good and it is unlikely to observe a great improvement with new models, then it is worthwhile considering a new set of features. Other reliability states, motivated by experiments described in (icassp06), are defined as follows.

$$\begin{array}{lll} \sigma_2 & : & \{g_A[SEG_a(b,e,t)] \neq g_G[SEG_a(b,e,t)]\} \\ & & \wedge\{KLD_a[SEG_a(b,e,t)] < 0.4\} \\ \sigma_3 & : & \{g_A[SEG_a(b,e,t)] = g_G[SEG_a(b,e,t)]\} \\ & & \wedge\{KLD_a[SEG_a(b,e,t)] > 0.4\} \\ \sigma_4 & : & \{g_A[SEG_a(b,e,t)] \neq g_G[SEG_a(b,e,t)]\} \\ & & \wedge\{KLD_a[SEG_a(b,e,t)] > 0.4\} \end{array} \tag{5}$$

Other states can be defined corresponding to other conditions whose characterization is motivated by experimental evidence.

3. DYNAMIC SELECTION OF FEATURE SETS

A strategy is now described for dynamically selecting a feature set during recognition.

Let $W = w_1 \dots w_h \dots w_H$ be the sequence of word and pause hypotheses generated by an initial decoder using features of $\Im^r$. Let $w_h = h_1 \dots h_k \dots h_K$ be the sequence of phonemes given in the lexicon of w_h.

Let us associate to each phoneme h_k a descriptor σ_k belonging to the alphabet of the variability/reliability states. Let $\sum_h = \sigma_1 \dots \sigma_k \dots \sigma_{K(h)}$ be the sequence of variability labels associated to the phonemes of w_h.

The features of $\Im_r$ are likely to be the cause of a wrong hypothesization of w_h if the probability $P[\overline{w}_h|\sum_h]$ is not low. The symbol $\overline{w}_h$ indicates the fact that hypothesis w_h is not correct. If the probability $P[\overline{w}_h|\sum_h]$ is above a given threshold for one word or for time segment containing a sequence of words and pauses, then a new set of features is computed in that segment and recognition is also performed with the new set of features. The experimental results described below have been obtained by using feature vectors $Y^m(nT)$ as reference and feature vectors $Y^j(nT)$ as additional features used only when the reference ones are not reliable.

The following approximation is proposed for computing $P[\overline{w}_h|\sum_h]$:

$$P[\overline{w}_h|\sum_h] = \frac{1}{K(h)} \sum_{k=1}^{K(h)} P(\overline{h}_k|\sigma_k) \tag{6}$$

where $P(\overline{h}_k|\sigma_k)$ indicates the probability that phoneme hypothesis h_k is incorrect. Such a probability is computed as follows:

$$P(\overline{h}_k|\sigma_k) = \frac{1}{1 + \frac{P(\sigma_k|h_k)}{P(\sigma_k|\overline{h}_k)} \frac{P(h_k)}{P(\overline{h}_k)}} \tag{7}$$

which is computed by cumulating counts for each phoneme in all contexts.

Other specific thresholds are used for dealing with the cases of word insertion and deletion.

As the ASR systems use acoustic models which are trained with a general telephone corpus without using any data of the application which is being tested, the training sets of AURORA3 have been used for tuning the just describes strategy.

When word hypotheses are generated with feature vectors $Y^j(nT)$, it is possible that there is a word consensus with hypotheses generated with the reference features. It is possible to investigate whether or not there is a consensus on an error. This has not been done yet and the problem will be investigated in future work. In the absence of word consensus, a decision is made based on the hypothesis with the lowest probability of being wrong. Notice that AURORA3 is a corpus of connected digits in strings of variable and unknown length. The language models is just a simple finite state automaton with word models in parallel an feed-back from the final to the initial state.

4. EXPERIMENTS

Experiments have been performed with the Italian and Spanish components of the Aurora3 database (connected digits collected in car environment). The acoustic models employed were hybrid HMM-NN trained on large corpora completely disjoint from Aurora3 namely the domain independent, phonetically balanced SpeechDat1-2 corpora.

The training corpora are made of telephonic read speech and were recorded in quiet environments. Different HMM-ANN models were trained, one for J-RASTA PLP and one for MRA, with the same training set for each language.

The Aurora3 corpus contains a set of close-talking utterances indicated as CH0 and a set of hand-free utterances, indicated as CH1. Utterances of CH0 are nearly clean, as the close-talking microphone collects little environmental noise, while utterances of CH1 are quite noisy as the hand-free microphone gathers a lot of car noise. Aurora3 is divided into training and test components.

The test corpus was used for producing the results, in terms of WER, reported in Table 1. Baseline results were obtained with MRA features which resulted to perform better than JRASTAPLP features for this task and with this setting (Gemello). *Oracle* results refer to what is obtained by comparing MRA and JRASTAPLP result with the reference and always deciding for the correct result if it is produced by at least one of the systems. *This strategy* results are the results obtained with the strategy proposed in this paper.

corpus	baseline	this strategy	oracle
Italian CH1	21.13	17.66	15.03
Spanish CH1	12.3	8.68	6.8

Table 1: Results in terms of WER with the baseline, the Oracle and the strategy proposed in this paper

With the proposed procedure, 339 (54%) sequences for the Italian and 309 (50.4%) for the Spanish where validated with the (6). The WER on the validated sequences were 4.8% for the Italian and 0.5% for Spanish. The percentage of words evaluated with the two feature streams for which there was consensus among the two recognizers was 11.5% for Italian and 18.3% for Spanish. Among these words, the WER was 33.7% for Italian and 12.8% for Spanish. The results show differences between the two languages, but a similar trend. The new feature set is applicable and provides different results in less than half of the word hypotheses. When it is applicable and the two recognizers do not provide the same result, the strategy leads to a significant WER reduction. The WER reduction is 16.42% for the Italian and 29.4% for Spanish.

5. CONCLUSIONS

A general approach for integrating different acoustic feature sets and acoustic models has been presented. A strategy for using a feature set as a reference and for scheduling the execution of other feature sets has been introduced. Additional features are computed and used only in case of high probability that the initial feature set has generated an incorrect word hypothesis. Experiments have been conducted using ANN and GMM acoustic models and features obtained with MRA and JRASTAPLP. Significant WER reductions have been observed on the test sets of the AURORA3 corpus.

6. ACKNOWLEDMENTS

This work was supported by European Commission 6th Framework Program project DIVINES under the contract number FP6-002034.

REFERENCES

[1] L. Barrault, D. Matrouf, R. De Mori, R. Gemello and F. Mana, “Characterizing Feature Variability in Automatic Speech Recognition Systems”, in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Toulouse France, 2006.

[2] H. Bourlard and S. Dupont, “Sub-band based speech recognition” , in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Munich Germany, 1997, pp. 1251-1254.

[3] R. Gemello, F. Mana, D. Albesano and R. De Mori, “Multiple resolution analysis for robust automatic speech recognition”, in *Computer Speech and Language*, 20(1), pp. 2-21, 2006.

[4] R. M. Hegde, H. A. Murthy and G. V. R. Rao, “Speech processing using joint features derived from the modified Group delay function”, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, PA, 2005, pp.I541-544.

[5] M. Kamal Omar and M. Hasegawa-Johnson, “Maximum Mutual Information Based Acoustic-Features Representation of Phonlogical Features For Speech Recognition”, in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Orlando, FL, 2002, pp. I 81-84j.

[6] D. Pearce, and H. Hirsch, “The Aurora experimental framework for the performance evaluation of speech recognition for mobile application”, in *International Conference on Spoken Language Processing*, Beijin, China, 2000, pp. 29-32.

[7] P. Pujol, S. Pol, C. Nadeu, A. Hagen, and H. Bourlard, “Comparison and combination of features in a hybrid HMM/MLP and a HMM/GMM speech recognition system”, in *IEEE Transactions on Speech and Audio Processing*, 2005, SAP-13(1):14-22.

[8] R. Sarikaya, Y. Gao, M. Picheny and H. Erdogan, “Semantic Confidence Measurement for Spoken Dialog Systems”, in *IEEE Transactions on Speech and Audio Processing*, 2005, SAP-13 (4) : 534-545.

[9] C. E. Shannon, “A Mathematical Theory of Communication ” in *The Bell System Technical Journal*, July, October 1948, Vol. 27, pp. 379-423, 623-656.

[10] O. Siohan, B. Ramabhadran and B. Kingsbury, “Constructing ensembles of asr systems using randomized decision trees”, in *IEEE Intl. Conference on Acoustics, Speech and Signal Processing*, Philadelphia, PA, March 2005, I, pp. 197-200.

[11] D.L. Thomson and R. Chengalvarayan (1998) “Use of periodicity and jitter as speech recognition feature”, in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, WA, May 1998, vol. 1, pp. 21 - 24.

[12] S.V. Vaseghi, N. Harte and B. Miller, “Multi resolution phonetic/segmental features and models for HMM-based speech recognition”, in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Munich Germany, 1997, pp. 1263-1266.

[13] A. Zolnay, R. Schluter and H. Ney, “Robust speech recognition using a voiced-unvoiced feature”, in *Proc. Int. Conf. on Spoken Language Processing*, Denver, CO, vol. 2, pp. 1065 - 1068.

[14] A. Zolnay, R. Schluter and H. Ney, “Acoustic feature combination for robust speech recognition”, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, PA, March 2005, pp. I 457-460.