A GENERAL METHOD FOR COMBINING ACOUSTIC FEATURES IN AN AUTOMATIC SPEECH RECOGNITION SYSTEM

Driss Matrouf, Loic Barrault, Renato De Mori

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

ABSTRACT

A general method for the use of different types of features in Automatic Speech Recognition (ASR) systems is presented. A gaussian mixture model (GMM) is obtained in a reference acoustic space. A specific feature combination or selection is associated to each gaussian of the mixture and used for computing symbol posterior probabilities. Symbols can refer to phonemes, phonemes in context or states of a Hidden Markov Model (HMM). Experimental results are presented of applications to phoneme and word rescoring after verification. Two corpora were used, one with small vocabularies in Italian and Spanish and one with very large vocabulary in French.

1. INTRODUCTION

It is known that Automatic Speech Recognition (ASR) systems make errors (see, for example, [17]). This is due to the imperfection of the various models used, on the limitations of the feature extracted and on the apporximations 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. 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 ([22], [12], [8], [9], [6], [4], [11], [13], [14], [24], [16], [5]).

At another level, attempts have been recently reported ([23], [21]) on the use of neural networks, decision trees and other machine learning techniques to combine the results of ASR systems fed by different feature streams or using different models in order to reduce word error rates (WER). In ([26]) it is shown that log-linear combination provides good results when used for integrating probabilities provided by acoustic models.

Other approaches integrate some specific parameters into a single stream of features ([20], [25], [19]). 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 ([15]), 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.

An analysis of the factors affecting WER in ASR systems is ceretainly useful for an effective use of differentr feature sets. In ([10]) ASR errors of spoken dialog data collected from various telephony-based customer care services are analyzed. It is shown that a combination of utterance-based Signal-to-Noise Ratio (SNR) and its local variations provide useful predictions of recognition error rate. The correlation between SNR and WER is also made evident in the study described in ([18]) regarding the English portion of the MALACH corpus for which an influence on WER of the number of syllables per second is also reported. In ([7]) the importance of correctly defining syllable boundaries is dis-

cussed and it is shown that ASR errors for certain consonants depend on whether a consonant is in a prevocalic or a postvocalic position in a syllable.

The approach described in this paper considers 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:

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

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. INTEGRATING 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:

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

Where $w_\mu^a[Y^r(nT)]$ are weights depending on the feature sample in a reference space indicated by the superscript 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.

Features $Y^r(nT)$ may be reliable in certain zones of the acoustic space $\Im^r$ and less reliable in other zones. Furthermore, in certain zones different models may provide better probability approximations than others.

The choice and use of models and features in a given point of $\Im^r$ depends on the values of the weights $w_\mu^a[Y^r(nT)]$. Let $g(nT) = w_\mu^a[Y^r(nT)]$ be a vector of weights corresponding to a point of $\Im^r$.

In practice, only a limited number of vectors of weights can be considered in a system. Let G be the set of these vectors. If only one feature stream is available with symbol posterior probabilities computed with two models, for example Artificial Neural Networks (ANN) and Gaussian Mixture Models (GMM) and the symbols are states of a Hidden Markov Model (HMM) then the vectors contain the coefficients of a log-linear combination of the probabilities provided by the two models. If vector elements are binary variables, and only one of them is equal to one, then changing vectors corresponds to switch between a feature stream and another.

In practice, vector values can be estimated or determined by experiments only in certain points of the reference space $\Im^r$ and the symbol posterior probability in a point of the acoustic space has to be estimated by smoothing the probabilities estimated with different vectors of weight values. For this purpose, a probability density $P[g|Y^r(nT)]$ is introduced. It indicates the probability that vector $g \in G$ provides the right weight model for computing $\log P(g|Y_n)$ with the (1). Different weight models for computing a symbol posterior probability are then used as follows:

$$P(q|Y_n) = \sum_{g \in G} P[qg|Y_n] = \sum_{g \in G} P_g[q|Y_n] P[g|Y^r(nT)] \quad (2)$$

where $P_g[q|Y_n]$ indicates the posterior symbol probability is computed with the (1), using the weight model g.

Probability $P[g|Y^r(nT)]$ can be computed by associating to each weight model g a Gaussian distribution as follows:

$$P[g|Y^r(nT)] = \frac{N(\mu_g, \Sigma_g, Y^r(nT))}{\sum_{\gamma \in G} N(\mu_\gamma, \Sigma_\gamma, Y^r(nT))} \quad (3)$$

There are many possible uses of the (1) and the (2). An initial attempt has been made by considering two sets of features, one of which is used as reference. Zones of

the acoustic space $\Im^r$ where the reference set of features is unreliable are determined and a new set of features is used in those zones. Hypothesis generation in these zones depends on the two feature streams. The (2) is used, in this case, as follows. A set of Gaussian distributions $P[g|Y^r(nT)]$ is obtained in the reference space. In each point Y^r of $\Im^r$ there is a Gaussian distribution $N(Y^r)$ with the highest value. If features in $\Im^r$ are reliable, then the features in reference space $\Im^r$ are associated to $N(Y^r)$, i.e. $g(Y^r) = \{\text{features in } \Im^r\}$, otherwise the new set of features is associated to the point. Other solutions are possible by associating to $g(Y^r)$ a log linear combination of models and/or features according to the (1).

A method for predicting the reliability of features based on their variability in a point Y^r is described in ([1]).

3. EXPERIMENTAL RESULTS ON PHONEME RECOGNITION

Two experiments have been conducted on the use of the (2) for computing posterior phoneme probabilities. They have been conducted on the French ESTHER corpus [2] and on the Italian portion of the CH1 part of AURORA3. The French ESTHER corpus has a very large vocabulary and contains 80000 broadcast news sentences in French. A set of 10000 sentences have been isolated to infer a mixture of 1024 distribution probabilities $P[g|Y^r(nT)$ in a reference space of PLP features. With Maximum A Posteriori (MAP) probability estimation, a mixture of 1024 Gaussians has been obtained for each of the 37 phonemes in French. After elimination of the Gaussians which did not have samples associated to them, a total of 10000 Gaussians were kept. A portion of the corpus consisting of about 60000 sentences was used to estimate a vector $P_g[q|Y_n]$ associated to each Gaussian g of the mixture. This was obtained by introducing a counter per phoneme and incrementing, after forced alignment and for each time frame, the counter of the phoneme f corresponding to a segment by a quantity equal to $P_g[f|Y_n]$. The contents of phoneme counters have been normalized to ensure that the same number of frames is used for each phoneme.

A test set containing 20000 phonemes was used. After forced alignment, classification was performed on each segment, ignoring the phonemes used for the alignment. In each segment, the (2) was applied and the phoneme with the highest posterior probability was considered as the recognition result. Such a result was compared with the one obtained using an HMM model per phoneme.

Notice that probabilities $P_g[f|Y_n]$ can be computed once forever using different features for different Gaussians g .

A similar test was conducted with the CH1 Italian portion of AURORA3. The training set was used for obtaining the GMM in the reference space and log-linear interpolation was used for computing $P_g[q|Y_n]$. The results are shown in Table 1.

corpus	using the (2)	using HMMs
Italian CH1	9.2	14
French ESTER	30.5	37.7

Table 1. Results in terms of phoneme error rates(%) using the (2) and phoneme HMMs

With a KLD lower than 0.004, then 63% of the Italian digits are validated with a phoneme error rate of 0.59%.

4. HYPOTHESIS VERIFICATION AND FEATURE SWITCHING

A simple application of the (1) is for hypothesis verification. The features in $\Im^r$ are initially used for generating word hypotheses.

An hybrid system consisting of an Artificial Neural Network and a set of Hidden Markov acoustic Models (HMM) is used. The recognizer uses a feature set obtained by Multi Resolution Analysis (MRA) followed by Principal Component Analysis (PCA). A denoising technique described in ([3]) is used. The stream of acoustic features will be indicated as $\{Y^m(nT)\}$. The value of T is 10 msecs and the feature set contains seven analysis frames centered on the frame at nT. Let $\Im^m$ be the space of the MRA features. In this case $\Im^r = \Im^m$.

The ANN is trained to recognize phonemes and transitions using a corpus of phonetically balanced sentences which are completely independent from the test data. The ANNs have 636 outputs, one for each phoneme and each transition between two successive phonemes.

A new set of features is obtained with Perceptual Linear Prediction (PLP) followed by RASTA filtering.

These features will be called JRASTAPLP and the corresponding vector will be indicated as $\{Y^j(nT)\}$ belonging to the acoustic space $\Im^j$. The vectors $Y^m(nT)$ and $Y^j(nT)$ represent two different observations of a speech segment Y_n centered on the same sample.

The two ASR systems separately fed by the two feature streams use acoustic models which are trained with a general telephone corpus without using any data of the application which is being tested. The ANNs of the two systems have the same topology and he same denoising algorithm is applied to the two feature streams.

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 $w_h = h_1 \dots h_k \dots H_{K(h)}$ be the sequence of phonemes given in the lexicon of w_h.

Initially only feature streams $Y^m(nT)$ are used for recognition, but the two feature streams are used for verification. The hypothesis w_h is generated in the time interval (t_{hb}, t_{he}), with segments labelled with phoneme hypotheses. Let assume that phoneme h_k is hypothesized in the time interval (t_{h_kb}, t_{h_ke}).

Two posterior probability streams $P_G^m[q|SEG(h_kb, h_ke)]$ and $P_G^j[q|SEG(h_kb, h_ke)]$ are computed in each segment $SEG(h_kb, h_ke)$. The probabilities are computed using segment Gaussian Mixture Models. The Kullback-Leibler distance (KLD) between the two streams is then computed:

$$KLD[SEG(h_kb, h_ke)] = D\left[P_G^m[q|SEG(h_kb, h_ke)]||P_G^j[q|SEG(h_kb, h_ke)]\right] = \sum_{q\in Q} P_G^m[q|SEG(h_kb, h_ke)] \log \frac{P_G^m[q|SEG(h_kb,h_ke)]}{P_G^j[q|SEG(h_kb,h_ke)]} \quad (4)$$

The features of $\Im^r$ are likely to be the cause of a wrong hypothesization of w_h if the probability $P[\overline{w}_h|Y_n(t_{hb}, t_{he})]$ is not low.

The symbol $\overline{w}_h$ indicates the fact that hypothesis w_h is not correct. If the probability $P[\overline{w}_h|Y_n(t_{hb}, t_{he})]$ is above a given threshold for one word or for a time segment containing a sequence of words and pauses, then the set of features $Y^j(nT)$ is considered in that segment and recognition is also performed using feature vectors $P_g(q|Y_n)$ in the (1) with a new selection g.

The following approximation is proposed for computing $P[\overline{w}_h|Y_n(t_{hb}, t_{he})]$:

$$P[\overline{w}_h|\sigma_{kld}] = \frac{1}{H(K_h)} \sum_{k=1}^{H(K_h} P(\overline{h}_k|\sigma_{kld}) \quad (5)$$

where $P(\overline{h}_k|\sigma_{kld}) = 1 - P(h_k|\sigma_{kld})$
and $\sigma_{kld} = KLD[SEG(h_kb, h_ke)]$.

The hypothesis w_h is accepted if $P[\overline{w}_h|Y_n(t_{hb}, t_{he})]$ is below a given threshold. Other specific thresholds are used for dealing with the cases of word insertion and deletion.

It is possible that contiguous word hypotheses generated with features $\{Y^m(nT)\}$ are incorrect. In this case, the segment corresponding to the sequence is processed with the new feature set $\{Y^j(nT)\}$ leading to a sequence of word hypotheses with a different number of words.

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.

5. EXPERIMENTAL RESULTS ON VERIFICATION

Experiments have been performed with the Italian components of the Aurora3 database (connected digits collected in car environment). The acoustic models employed were hybrid HMM-ANN 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 2.

Baseline results were obtained with MRA features which resulted to perform better than JRASTAPLP features for this task and with this setting ([3]). *Oracle* re-

sults 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	17.04	15.03

Table 2. Results in terms of WER with the baseline, the Oracle and the strategy proposed in this paper

6. CONCLUSIONS AND FUTURE WORK

A general method for the combined use of different feature sets has been proposed. Specific symbol posterior probability computation is associated to each gaussian of a mixture. These probabilities can be computed once forever and simply retrieved during decoding or rescoring. Examples of application have been presented showing that the proposed method of computing phoneme posterior probabilities is effective even when a single feature set is used.

As an extension of the (2), it possible to consider two acoustic spaces as reference. In this case, symbol posterior probabilities are computed as follows:

$$P(q|Y_n) = \sum_{g_m,g_j \in G} \left\{ P_g[q|Y_n^m, Y_n^j] * P[g_m|Y_n^m] * P[g_j|Y_n^j] \right\} \quad (6)$$

7. ACKNOWLEDMENTS

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

8. 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] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.F. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of French broadcast news", in *Proc. European Conference on Speech Communication and Technology*, Eurospeech 05, 1149-1152.

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

[4] R. Hariharan, I. Kiss and O. Viikki , "Noise robust speech parameterization using multiresolution feature extraction", in *IEEE Transactions on Speech and Audio Processing*, 2001, SAP-9(8): 856-865.

[5] 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.

[6] H. Hermansky and N. Morgan, "RASTA Processing of Speech", in *IEEE Transactions on Speech and Audio Processing*, October 1994, Vol. 2, n 4, pp. 578-589.

[7] M.J. Hunt, "Speech recognition, sylabification and statistical phonetics", in *Proc. International Conference on Spoken Language Processing*, Interspeech 04, 57-60.

[8] H.W. Hon and K. Wang, "Combining frame and segment based models for large vocabulary continuous speech recognition", in *Proc. IEEE ASRU Workshop*, Keystone, Colorado, 1999.

[9] K. Jiang and X. Huang, "Acoustic feature selection using speech recognizers", in *Proc. IEEE ASRU Workshop*, Keystone, Colorado, 1999.

[10] H. K. Kim and M. Rahim, "Why Speech Recognizers Make Errors? A Robustness View", in *Proc. International Conference on Spoken Language Processing*, Jeju, Korea, 2004, ThA1703o1.

[11] B. Kingsbury, G. Saon, L. Mangu, M Padmanabhan and R. Sarikaya, "Robust speech recognition in noisy environments: The 2001 IBM spine evaluation system", in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Orlando, FL, 2002, pp. I 53 - 56.

[12] K. Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noise and reverberant environments", in *Proc. International Conference on Spoken Language Processing*, Sydney, AUS, 1998, pp.891-894.

[13] M Kleinschmidt and D. Gelbart, "Improving word accuracy with Gabor feature extraction", in *Proc. International Conference on Spoken Language Processing*, Denver, CO, 2002, pp. 25-28.

[14] N. Morgan, B. Chen, Q. Zhu and A. Stolcke, "TRAPping Conversational Speech: Extending TRAP/Tandem approaches to conversational telephone speech recognition", in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Montreal, Canada, 2004.

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

[16] 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.

[17] 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.

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

[19] T.A. Stephenson, M.M. Doss and H. Bourlard, "Speech recognition with auxiliary information", in *IEEE Transactions on Speech and Audio Processing*, 2004, SAP-12(3):189-203.

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

[21] T. Utsuro, Y. Kodama, T. Watanabe, H. Nishizaki and S. Nakagawa, "Confidence of agreement among multiple LVCSR models and model combination by SVM", in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, China, 2003, pp. I-16, 19.

[22] 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.

[23] R. Zhang and A.I. Rudnicky, "Word level confidence annotation using combinations of features", in *Proc. European Conference on Speech Communication and Technology*, Eurospeech 01, Aalborg, Denmark, 2001, pp. 2105-2108.

[24] Q. Zhu, B. Chen, N. Morgan and A. Stolcke, "On Using MLP Features in LVCSR", in *Proc. International Conference on Spoken Language Processing*, Jeju, Korea, 2004, paper WeA501p5.

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

[26] 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.