CHARACTERIZING FEATURE VARIABILITY IN AUTOMATIC SPEECH RECOGNITION SYSTEMS

Loic Barrault§, Driss Matrouf§, Renato De Mori§, Roberto Gemello, Franco Mana**

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

* LOQUENDO
Via Valdellatorre, 4 – 10149 Torino – Italy
roberto.gemello,franco.mana@loquendo.com

ABSTRACT

A method is described for predicting acoustic feature variability by analyzing the consensus and relative entropy of phoneme posterior probability distributions obtained with different acoustic models having the same type of observations. Variability prediction is used for diagnosis of automatic speech recognition (ASR) systems. When errors are likely to occur, different feature sets are considered for correcting recognition results.

Experimental results are provided on the CH1 Italian portion of AURORA3.

1. INTRODUCTION

Intrinsic feature variability depends on the set of classes that features have to discriminate. Most frequently considered classes are phonemes or phonetic features. Feature variability causes ambiguities in classifying speech signal segments. Ambiguities can be reduced by using different feature streams.

A single feature stream can be obtained from different streams. In [6], a sub-optimal solution is proposed for selecting features from two different sets. Other approaches integrate some specific parameters into a single stream of features [9].

Without attempting to find an optimal set of acoustic measurements, many recent automatic speech recognition (ASR) systems combine streams of different acoustic measurements ([4], [10]). In [11] it is shown that log-linear interpolation provides good results when used for integrating probabilities computed with acoustic models with different feature sets. Another possibility [8] consists in combining the results of ASR systems in order to reduce word error rates (WER).

In this paper, the possibility is considered to use multi-model systems for predicting feature variability as described in section 2. A solution is presented in section 3. Using variability indicators as diagnosis features, the possibility of predicting when it is potentially useful to use a new set of features is discussed in section 4. Results with a programmable use of a multi-feature system are presented in Section 5.

2. PROBLEM DESCRIPTION

A set *a* of acoustic features is used in an ASR front-end to segment the speech signal and assign to each segment hypotheses about class symbols $q \in Q$, where Q is an alphabet. Scores are assigned to the hypotheses. A frequently used score is the posterior probability $P\left[q \middle| Y^a(nT)\right]$ computed with acoustic models. $Y^a(nT)$ is a vector of values of the elements of a feature set $\Im^a$ identifying a point in the acoustic space Γ^a. T is the interval between two successive analysis frames. Symbols q may represent a phoneme, a phoneme in context, a transition between two phonemes or a state of a Hidden Markov Model (HMM). The impact of features on recognition results depends on many factors such as the number and complexity of the models and the frames used as observations. In spite of the use of context-dependent and speaker-dependent models, a tangible amount of errors remains which may depend on the imperfection of acoustic models or on intrinsic ambiguity of features. In the attempt to separate the effect of model imperfection from the effect of feature variability, the relative entropy of phoneme posterior probability distributions obtained with different models is considered. This relative entropy can be computed in a point of the acoustic space. When it is low, it is likely that the resulting equivocation in phoneme recognition is due to intrinsic feature variability. Furthermore, a predictor of equivocation is a useful element for the ASR system diagnosis. Equivocation between a channel source S which emits symbols $f \in Q$ and the receiver R which hypothesizes symbols $g \in Q$ is defined as follows:

$$H_R(S) = -\sum_{f,g} P\{f,g\} \log P\{g|f\} \qquad (1)$$

Notice that equivocation can be directly compared to the source entropy, while symbol error rates and vocabulary size or language perplexity are more difficult to compare since their dimensions are different.

The coverage of the acoustic space in which the equivocation is expected to be low for an application corpus provides an indication of the degree of success for the

1-4244-0469-X/06/$20.00 ©2006 IEEE

application. Thus, it is important to find diagnostic confidence measures capable of predicting the degree of equivocation in points of the acoustic space. For the values $Y^a(nT)$ for which equivocation is expected to be high, another set of features can be considered. The new set of features is likely to be useful if its relative entropy with respect to the initial set of features is high and if the expected equivocation in the corresponding points is low. An ASR setup with two different feature sets is considered in this paper together with two very different acoustic models, namely Artificial Neural Networks (ANN) and Gaussian Mixture Models (GMM). Furthermore, the different models are trained with different types of data in different conditions.

3. COHERENCE OF HYPOTHESES GENERATED WITH DIFFERENT MODELS

Posterior probabilities of phonemes given the acoustic observations are obtained with ANNs and GMMs. They are indicated as $P_{ANN}\left\{q \middle| Y^a(nT)\right\}$ and $P_{GMM}\left\{q \middle| Y^a(nT)\right\}$. The ANN is a time-delay Neural Network which computes the probability of being in a state of an HMM, given the observation made of a set of input frames. This hybrid HMM-ANN system is described in [1]. The input window is 7 frames wide, and each frame contains the set of features extracted by the front-end along with their first and second time derivatives. There are two hidden layers. The second hidden layer is fully connected with the output layer that estimates 686 emission probabilities of phonemes and diphone transitions. Only phoneme probabilities are considered in this study. The ANN parameters are trained with a rich corpus of generic telephone conversations. The GMMs are mixtures of 512 gaussians per phoneme. Their parameters are estimated with Maximum Likelihood (ML) estimation with the training set of each specific application and phoneme segments obtained with the ANN. They are introduced for deriving confidence indicators and not for use in an independent ASR system. The time segment in which acoustic features are computed is the same for both model types, but it could be different. In general, many model types could be considered by varying the phonetic context of the same phoneme or by varying the acoustic context in which parameters describing time frame nT are computed. Models with or without adaptation could also be compared.

The 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 relative entropy between the two posterior probability distributions $P^a_{ANN}(t) = P^a_{ANN}\left\{q \middle| SEG_a(b,e,t)\right\}$ and $P^a_{GMM}(t) = P^a_{GMM}\left\{q \middle| SEG_a(b,e,t)\right\}$, is the Kullback-Leibler distance (KLD) indicated as:

$$
\begin{aligned}
&KLD_a(t) = KLD_a\left[SEG_a(b,e,t)\right] = \\
&= D\left[P^a_{ANN}\left\{q \middle| SEG_a(b,e,t)\right\} \middle\| P^a_{GMM}\left\{q \middle| SEG_a(b,e,t)\right\}\right] = \quad (2) \\
&= \sum_{g \in Q} P^a_{ANN}\left\{g \middle| SEG_a(b,e,t)\right\} \log \frac{P^a_{ANN}\left\{g \middle| SEG_a(b,e,t)\right\}}{P^a_{GMM}\left\{g \middle| SEG_a(b,e,t)\right\}}
\end{aligned}
$$

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 $g^a_A(t) = g^a_A(b,e,t)$ and $g^a_G(t) = g^a_G(b,e,t)$.

In [7], the average relative entropy has been used for selecting a feature set in a group of potential candidates. Here, relative entropy is used for measuring the divergence of the outputs of two systems fed by the same input data. Posterior probabilities obtained with these models may be inaccurate. Inaccuracy is reduced by combining these posterior probabilities with log-linear interpolation as proposed in [11]:

$$P^a(t) = \alpha \log P^a_{ANN}(t) + (1-\alpha) \log P^a_{GMM}(t) \quad (3).$$

The interpolation coefficient α is determined to maximize the phoneme recognition rate on the training set. Indicators of model accuracy are $KLD_a(t)$, and the fact that different models assigns the maximum posterior probability to the same symbol. This is indicated by the truth of the predicate $c_a(t) = T \text{ iff } g^a_A(t) = g^a_G(t)$.

4. RELATION BETWEEN MODEL DIVERGENCE AND CHANNEL EQUIVOCATION

Two different types of feature streams are considered. The first set is based on Multi Resolution Analysis (MRA). Motivations for using these features and details are described in [2]. The other is based on 12 J-RASTA Perceptual Linear Prediction (PLP) coefficients [3] with their first and second time derivatives plus total energy and its time derivatives. An initial experiment was performed on the phonemes of the Italian portion of the CH1 (noisy) part of the AURORA3 corpus. The GMMs were trained using the Italian portion of the training corpus of AURORA3. The test set of the Italian portion of AURORA3 was used for both models. A channel model, represented in Figure 1, is considered for computing equivocation after forced alignment. A computation unit estimates, for all symbols, the posterior probabilities $P^m_A(t)$ and $P^m_G(t)$. The log-linear interpolation of them is then computed for hypothesizing the phoneme $g^m(t)$. $KLD_m(t)$ is considered as an indicator of confidence related to modeling difficulty for the segment $[SEG_m(b,e,t)]$.

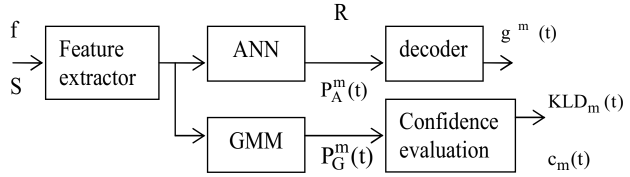

Figure 1- Diagnosis channel model

The source entropy is 3.5306 bits. Figure 2 shows the relation between equivocation and KLD(t) for the two feature sets. Along the X axis are values of X such that $KLD_m(t) < X$. MRA features appear to be more suitable for the application considered, the set of symbols and the models used. Nevertheless, the main difference appears for high values of KLD.

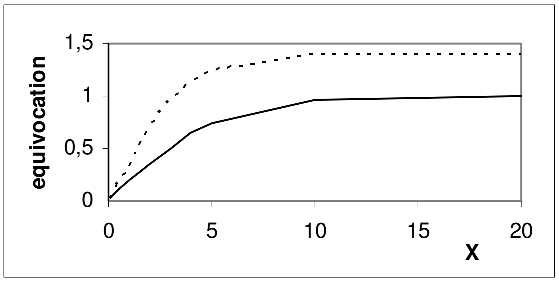

Figure 2 – Comparison of equivocations obtained with MRA features (continuous line) and JRASTAPLP features (dotted line)

The relation between the truth of $c_m(t)$ and equivocation is also worth considering. Figure 3 shows the relation between equivocation and $KLD_m(t)$ depending on the consensus expressed by the truth of $c_m(t)$. A maximum of equivocation of 0.221 with coverage of 54.11% is observed for all data when $c_m(t)$ is true. An overall equivocation of 0.11 was observed for $KLD_m(t) < 0.5$ with a coverage of 40.55%. The overall equivocation is computed with the log linear interpolation of the probabilities computed with the two models. Three states, representing increasing variability expectations, can be identified from these data corresponding to $KLD_m(t) < 0.5$ (VS1), $KLD_m(t) > 0.5$ and consensus (VS2) $KLD_m(t) > 0.5$ (VS3). Indicators of variability are useful confidence descriptors. Notice that they are not necessarily related to the entropy measures proposed in [5]. Similar relations are obtained with the train and test sets and with JRASTA PLP features. With JRASTA PLP features an equivocation of 0.17 is found for KLD(t)<0.5 with a coverage of 18.72%. This suggests that, at least with the models used and the application considered, MRA features are expected to exhibit lower variability for a larger portion of data. A similar behavior with higher equivocation was found for the Italian portion of SpeechDat, a large vocabulary, continuous speech telephone corpus. The source entropy for this corpus is 4.1 bits and the equivocation for $KLD_m(t) < 0.5$ is 0.5.

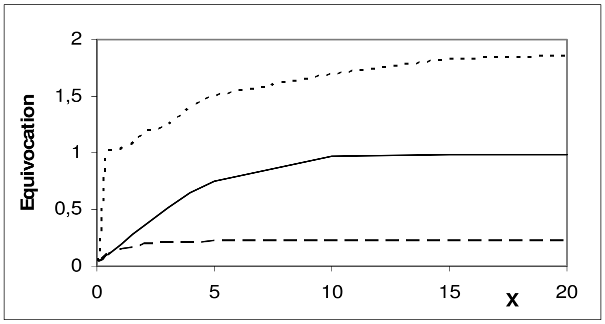

Figure 3 – Equivocation as a function of intervals of KLD(t) when MRA features are used with separate curves for the cases in which $c_m(t)$ is true (dashed) and false (dotted).

If $KLD_m(t)$ is very small (e.g. <0.1), it is very likely that equivocation is due to intrinsic feature variability. The corresponding equivocation depends on the complexity and environment of the application. For the noisy connected digits, an equivocation of 0.037 was found. This value becomes 0.4 for SpeechDat. As the equivocation measures depend only on the acoustic models and the features, improvements can be obtained by enriching alphabet Q with context dependent symbols and performing model adaptation. In spite of this, it is hard to reduce the equivocation by a factor of 10.

It is interesting to investigate the possibility of using JRASTAPLP features when variability with MRA features is in state VS3. For this purpose, let the vectors $Y_m(nT)$ and $Y_j(nT)$ respectively represent two different observations with MRA and JRASTAPLP features. Frame relative entropy is computed as follows:

$$KLD_{mj}[nT] = D\left[P_A^m(nT) \middle\| P_A^j(nT)\right] \qquad (4).$$

If $KLD_{mj}(nT)$ is low, that means that additional features do not provide a significant new amount of information. Notice that such a measure is independent from the lexicon and language models. The coverage as a function of KLD intervals has been analyzed. For KLD<1 a coverage of 88% was observed, indicating that the two feature sets provide very often rather similar probability distributions.

6. USING MULTIPLE FEATURE SETS FOR ASR

The possibility of predicting feature variability makes it possible to introduce a new paradigm for integrating different feature sets. Given a feature set, e.g. MRA, it is possible to estimate the parameters of a Gaussian mixture $\left\{\omega_j, N\left[\mu_j, \Sigma_j, Y^m(nT)\right]\right\}$. A partition in the space is obtained by considering zones in which each Gaussian $N\left[\mu_j, \Sigma_j, Y^m(nT)\right]$ provides the highest probability density. If the partition is detailed enough, one may assume that the posterior probability $P_j^m(q)$ of symbol q exhibits little variability in each zone. A posterior probability $P_j(q)$ can be estimated in a learning phase using the feature set or a combination of sets which provides the lowest equivocation in that zone. During recognition, posterior probabilities are computed as follows:

$$P(q|Y_n) = \sum_{j=1}^{J} \omega_j N\left[\mu_j, \Sigma_j, Y^m(nT)\right] P_j(q)$$

A simple experiment was performed using the CH1 portion of the Italian test set in AURORA3 by choosing, for computing $P_j(q)$, the most appropriate feature set between MRA and JRASTAPLP features. Phoneme posterior probabilities were computed with ANN.

The overall WER decreases from 20.34 with MRA features to 18.06% by switching feature sets for computing $P_j(q)$. It was also observed that 58% of the correctly hypothesized words with MRA/ANN exhibit consensus between all the phonemes in each word and the corresponding $g_G^m(t)$.

CONCLUSIONS

An approach to characterize feature variability has been proposed. The results are used to derive confidence indicators based on which the use a new feature set can be programmed. By using it on 37.5% of the sentences a WER reduction of 11.2 % was observed on the CH1 test set of the Italian portion of AURORA3.

New strategies will be investigated for performing a more accurate selection of speech segments for which high feature variability is expected. The possibility will also be investigated of introducing a programmable, more effective, local use of additional feature sets.

ACKNOWLEDMENTS

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

REFERENCES

1. R. Gemello, D. Albesano, F. Mana, "*Multi-source neural networks for speech recognition"*, in Proc. of International Joint Conference on Neural Networks (IJCNN'99), Washington, July 1999
2. R. Gemello, F. Mana, D. Albesano and R. De Mori "*Multiple resolution analysis for robust automatic speech recognition"*, Computer Speech and Language (accepted, in press).
3. H. Hermansky and N. Morgan, "*RASTA Processing of Speech"*, IEEE Transactions on Speech and Audio Processing, Vol. 2, n° 4, pp. 578-589, October. 1994.
4. M Kleinschmidt and D. Gelbart, "*Improving word accuracy with Gabor feature extraction"*, Proc. International Conference on Spoken Language Processing, Denver, CO, pp. 25-28, 2002.
5. H. Misra, H. Bourlard, and V. Tyagi *"New entropy based combination rules in HMM/ANN multi-stream ASR"* Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Hong Kong, China, pp. II-741–II-744, 2003.
6. M. Kamal Omar, Mark Hasegawa-Johnson "*Maximum mutual information based acoustic-features representation of phonological features for speech recognition"* Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Orlando, FL, pp. I 81-84, 2002.
7. M.K. Omar, K. Chen, M. Hasegawa-Johnson and Y. Bradman "*An evaluation on using mutual information for selection of acoustic features representation of phonemes for speech recognition*", Proc. International Conference on Spoken Language Processing, Denver, CO, pp.2129-2132, 2002.
8. O. Siohan B. Ramabhadran and B. Kingsbury "*Constructing ensembles of ASR systems using randomized decision trees"*, IEEE Intl. Conference on Acoustics, Speech and Signal Processing, Philadelphia, PA, March 2005, I, pp. 197-200.
9. D.L. Thomson and R. Chengalvarayan (1998) "*Use of periodicity and jitter as speech recognition feature"*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Seattle, WA, May 1998, vol. 1, pp. 21 – 24.
10. S.V. Vaseghi, N. Harte and B. Miller, "*Multi resolution phonetic/segmental features and models for HMM-based speech recognition"*, Proc. International Conference on Acoustics, Speech and Signal Processing, Munich Germany, pp. 1263-1266, 1997.
11. Zolnay, R. Schluter, and H. Ney, "*Acoustic feature combination for robust speech recognition",* Proc. ICASSP 2005, Philadelphia, PA, pp. I 457-460.