VARIABILITY OF AUTOMATIC SPEECH RECOGNITION SYSTEMS USING DIFFERENT FEATURES

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

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

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

ABSTRACT

The paper describes the use of two recognizers fed by different acoustic features. The first recognizer performs Multiple Resolution Analysis (MRA) while the other recognizer computes JRASTA Perceptual Linear Prediction Coefficients (JRASTAPLP). The two recognizers use the same denoising method but perform different partitions of their acoustic spaces. Experiments with the Italian and Spanish components of the AURORA3 corpus show that the two systems provide, in a significant proportion of cases, substantially different posterior probabilities for the same phoneme in the same time interval. A decision rule is proposed when two different words are hypothesized by the two recognizers. It is based on the probability that a hypothesis is correct, given the identity of the word hypotheses that are in competition. Significant word error rate (WER) reductions have been found for the CH1 proportion of the Italian and Spanish components of the AURORA3 corpus.

1. INTRODUCTION

Attempts have been recently reported [1][2] on the use of neural networks, decision trees and other machine learning techniques to combine the results of Automatic Speech Recognition (ASR) systems in order to reduce word error rates (WER). In [3], log-linear model combination and feature combination models are proposed to enhance ASR performance. In [4] a Bayesian model combination for ASR outputs is proposed. It computes the likelihood that a sentence hypothesized by each system is correct given system hypotheses and their confidence scores. Independence is assumed among systems and correctness probabilities depend on the overall system performance without considering which phonemes of words are hypothesized by each system. . Other recent work focuses on factors affecting WER in ASR systems. In [5], it is shown that a combination of utterance-based Signal-to-Noise Ratio (SNR) and its local variations provide useful predictions of recognition error rates.

The objective of the research described in this paper is to understand when and how different sets of features affect recognition performance. Specific probabilities are introduced that a phoneme hypothesis is correct given the phonemes that would be hypothesized with two different feature sets and their posterior probabilities. In this way, correctness probabilities explicitly depend on the competing word and phoneme hypotheses generated by systems using different feature sets. Two versions of an ANN/HMM hybrid recognizer 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) described in [6] and by Perceptual Linear Prediction (PLP) described in [7] followed by RASTA filtering. The latter features will be indicated as JRASTAPLP. The same denoising technique, described in [6] is used for both feature sets. 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.

Experiments were carried out with the test sets of the Italian and Spanish portions of the AURORA3 corpus. Intervals of phoneme posterior probabilities computed by the two systems are defined in section 2. Statistics of joint values of posterior probabilities for each phoneme class are reported showing important differences in the behavior of the two systems. An analysis of the consensus among the recognizers is presented in section 3. The probability that they generate correct hypotheses is high when their outputs are the same. A specific decision strategy is proposed in section 4 for the situation in which there is a discrepancy between the words hypothesized by the recognizers. Experimental results with this strategy are reported in section 5. Unfortunately, even if this is unlikely, different recognizers may generate the same wrong hypothesis. Diagnosis in this case is difficult, because situations of this type are rare and the output discrepancies cannot be used to guide further analysis. This aspect is not treated in this paper.

2. PERFORMANCE COMPARISON OF THE FEATURE SETS

The same sampled input signal S={s(nτ)}, where τ is the sampling period, is transformed by the two recognizers into two streams of acoustic features, namely $Y_m(nT)$ and $Y_j(nT)$ where T is the interval between two successive analysis frames and indices *m* and *j* respectively refer to MRA features and JRASTAPLP features. From now on, these two indices will be used to indicate the two types of features. The vectors $Y_m(nT)$ and $Y_j(nT)$ represent two different observations of a speech segment centered on the same sample. 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 Γ_m and Γ_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. Posterior probabilities for phonemes and diphones in a point of an acoustic space represent the *variability* of the features extracted. 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 phonemes *f*. Let $P\left\{f\middle|Y_m(nT)\right\}$ be the posterior probability of phoneme *f* given the MRA features at time (nT) and $P\left\{f\middle|Y_j(nT)\right\}$ be the posterior probability of phoneme *f* given the JRASTAPLP features at time (nT). Let us consider the space containing points having coordinates $P\{f|Y_m(nT)\}$ and $P\{f|Y_j(nT)\}$. Such a space can be partitioned as in Figure 1. Counts for the occupation of zones in this space can be collected for each pair of phonemes $\{f_m, f_j\}$. Each unit of the count represents the posterior probabilities generated by the two recognizers for the two phonemes when the same time frame is considered.

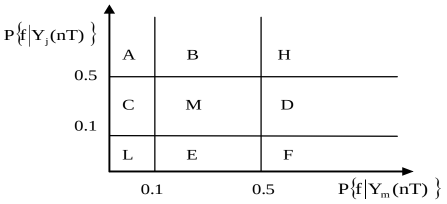

Figure 1. – Partition of the space of posterior probabilities.

In order to investigate feature variability for each phoneme and each set of features, the case $\{f_m = f_j\}$ has been considered and statistics have been collected for each phoneme *f*. This has been obtained by performing forced alignment of data in the AURORA3 training set (not used to train the ANNs) and considering the intersection of the time intervals hypothesized by the two systems for the same phoneme.

Zone H represents the case in which a phoneme is the most likely hypothesis with both feature sets. In this case the phoneme should be correctly hypothesized. Zone L correspond to the case in which hypothesizing the phoneme with both feature sets is similar to doing it with a random choice among candidates with uniform distribution, Zone M represents the case in which features indicate a possibility for the phoneme but that a decision would not be very reliable. Zones A, B and C indicate that JRASTAPLP better represents a phoneme than MRA and zones D, E and F indicate the contrary. Statistics collected about the fact that probabilities for a phoneme are represented by points in the above defined zones, may indicate possible confusions due to the inadequacy of the features or of the models to represent a phoneme. Moreover, other useful diagnostic results could be obtained when different modeling techniques or acoustic models in different languages are used.

An experiment was carried out with the training sets of the CH0 and CH1 data of the Spanish and Italian components of the AURORA3 corpus. The results, grouped by classes of phonemes, are reported in Table I. After forced alignment, phoneme segment hypotheses are generated for each stream of features. The average posterior probabilities, for each segment and for each feature set, have been computed and the pair was represented by a label in Figure 1.

Results clearly show that nonsonorant phonemes are more difficult to recognize than sonorant ones especially because plosives are short in time and features for fricatives are often distorted by denoising. Vowels are less affected by denoising because, in general, their segments have a fairly high SNR. MRA appears to provide better discrimination for many phonemes, but is particularly weak on semivowels. The reason is that these phonemes are often very short and a fine frequency resolution is required for their distinction. An analysis frame of long duration is required for obtaining the necessary frequency resolution with MRA. The average frame SNR after denoising varies between 6.5 and 14.8 for vowels, 3.6 and 10.4 for nonsonorant consonants, 6.0 and 10.2 for sonorant consonants.

The most impressive result is that there are many cases in which probabilities with one set of features are high while they are low with the other set of features. In the lines labeled with DEF, MRA features show better performance, while in the lines labeled with ABC, the JRASTAPLP features show better performance. It is rather unlikely that this result is only due to limits of the modeling techniques especially because there are many cases in which one probability is above 0.5 and the other is below 0.1. Likely, part of the discrepancies (cases A,B,D,F) are due to intrinsic feature variability which makes it difficult to infer appropriate probability distributions leading to good discrimination.

Table I – Distributions of the probability zones for the CH1 data of the Spanish and Italian components of the AURORA3 corpus (comma represents union)

	Language	Non sonorant	Sonorant consonant	Vowels
D,E,F	SPANISH	20,74%	13,39%	23,12%
	ITALIAN	20,49%	18,14%	12,59%
A,B,C	SPANISH	15,05%	13,8%	5,3%
	ITALIAN	16,01%	11,09%)	14,85%
H	SPANISH	40,36%	63,94%	66,4%
	ITALIAN	37,54%	61,04%	58,02%
M	SPANISH	2,31%	1,79%	1,53%
	ITALIAN	8,29%	3,87%	7,01%
L	SPANISH	21,55%	7,08%	3,64%
	ITALIAN	17,67%	5,86%	7,54%

The following remarks are also worth to be mentioned:

- the ANNs take into account a fairly long acoustic context,
- experiments were performed on a noisy small vocabulary corpus, thus phoneme recognition refer to limited phonetic contexts,
- ANNs were not trained on the application considered for the experiment nor on car data, rather they were trained on standard telephone environments,
- in general, the Italian corpus shows lower performance, especially on vowels in spite of the fact that the average SNR is greater than 10 after denoising for all the vowels but /uh/,
- the number of cases in D,E,F is generally higher than that in A,B,C showing that it is more frequent that the most likely phoneme with MRA is correct.

3. SITUATION DEFINITION

A sequence of words W generates a path in Γ_m and a path in Γ_j. The two paths are related by the fact that they are two representations of the same signal. The two recognizers label each path with a word sequence, namely W_m and W_j. Several situations can be defined based on the types of consensus between the words in W_m and W_j. The two sequences may be the same or they may differ by a variable number of words. In the latter case, it may be useful to identify the time segments in which the difference appears and analyze the types of discrepancies between the two recognizer outputs.

An initial set of situations or *result comparison states* are defined as follows:

- Q1: $W_m = W_j$;
- Q2: the same word or two different words are hypothesized in approximately the same time interval even if $W_m \neq W_j$, i.e.:

$$\exists(a,b)/\{w_{ma}=w_{jb}\}\wedge\{S(w_{ma})\approx S(w_{jb})\} \tag{1}$$

where $\{w_{ma} \in W_m\}$, $\{w_{jb} \in W_j\}$ and *S(x)* refers to the segmentation of *x*.

- Q3: two segments of W_m and W_j have approximately the same time bounds but without any word in common.

When the state of the recognition results is Q1, combining the scores of hypotheses generated by the two recognizers would produce the same result. It will be shown that in Q1 the WER is low. Diagnosis reveals that deletion errors in Q1 are essentially due to the imprecision of voice activity detection or denoising and not to the capability of the feature sets to discriminate among phonemes. If a portion of a word signal is considered as a non-speech segment, then the remaining part is often attached to a neighbor word which is likely to be misrecognized. Segmentation errors are a frequent cause of errors in Q1. Insertions are often due to the fact that background noise is considered as a speech segment.

Using the test sets of the CH1 data of the Spanish and Italian components of the AURORA3 corpus, in the case of full sentence consensus, the coverage is 72.66% for Spanish with a WER of 0.16% and 63.16% for Italian with a WER of 2%.

For the cases corresponding to state Q2, the WER is also low. Errors are essentially caused by substitutions. Some errors are due to segmentation and denoising but others reveal poor discrimination power among phonemes. This weakness is common to both feature sets in certain zones of their acoustic spaces, especially those corresponding to low SNR.

It is unlikely that correction of errors in the case of word consensus can be obtained with better strategies and scoring methods. Some errors can be avoided by the use of good language and lexical models. This aspect is not considered in this paper whose main objective is to report on the comparative study of feature performance.

In the absence of consensus, the oracle WER reveals that many errors could potentially be avoided by a scoring and decision strategy conceived to maximize the probability of selecting the correct candidate when it is proposed by only one of the two recognizers.

4. TYPES OF DIVERGENCE BETWEEN SYSTEM OUTPUTS

When there is no consensus among the hypotheses generated by the two recognizers, then their most likely hypotheses are aligned. Experiments described in detail in [6] show that the MRA recognizer has better performance in both languages than the other recognizer. So its hypotheses are considered as reference for aligning the results of the two recognizers when they disagree. Let $W_m(b,e)$ be a word or a sequence of words hypothesized by the MRA recognizer in the time interval (b,e) and $S_j(b,e)$ be the competing sequence of phonemes hypothesized by the JRASTAPLP recognizer in a time segment with substantial overlapping (more than 50%) with (b,e).

The situations defined in Table II, describing discrepancies among the recognizers, are worth considering.

Table II – Types of discrepancies between the outputs of different recognizers

MRA	RPLP	TYPE
$W_m(b,e):w_i$	$S_j(b,e):w_k$	*substitution* (sb)
$W_m(b,e):w_i$	$S_j(b,e):w_i w_k$	*j-insertion* (i_j)
$W_m(b,e):w_i w_k$	$S_j(b,e):w_i$	*m-insertion* (i_m)
$W_m(b,e):w_i$	$S_j(b,e):w_q w_k$	*j-substitution* and insertion (si_j)
$W_m(b,e):w_q w_k$	$S_j(b,e):w_i$	*m-substitution* and insertion (si_m)
All the other cases		multiple discrepancies(md_{mj})

The following decision strategy is proposed when there is no sentence consensus. The MRA word candidate is selected, except for situation *sb* for which the decision rule introduced below is applied.

Let $C(w, w_m, w_j)$ represents the fact that *w* is correct when the hypotheses with the highest score obtained with the MRA and the JRASTAPLP feature sets are respectively w_m and w_j. The decision rule is:

$$\begin{aligned} w^* &= \operatorname*{arg\,max}_{w\in(w_m,w_j)} P\{C(w)|w_m,w_j,\sigma\} = \\ &= \operatorname*{arg\,max}_{w\in(w_m,w_j)} P\{\sigma|C(w,w_m,w_j)\}P\{C(w,w_m,w_j)\} \end{aligned} \tag{2}$$

where C(w) is a predicate which is true when hypothesis *w* is correct and $\sigma:[\sigma_1(w),..,\sigma_n(w),..,\sigma_N(w)]$ represents a sequence of labels defined in Figure 1. As the hypothesis *w* is available with its segmentation, each segment of *w* corresponds to a phoneme, so that *w* can be represented by the sequence of phonemes $w:[f_1(w),..,f_n(w),..,f_N(w)]$.

For every phoneme $f_n(w)$, it is possible to consider the phoneme hypothesized by the other recognizer which has the highest number of frames in common with $f_n(w)$. The posterior probabilities of the two phonemes are represented by a symbol $\sigma_n(w)$ according to the grid defined in Figure 1.

Probability $P\{C(w, w_m, w_j)\}$ is computed from the training set which has not been used to train the acoustic models. For large vocabularies, this probability can be obtained as a product of prior probabilities of syllables or even of phonemes. Notice that the probability that neither w_m nor w_j is correct can be obtained by subtracting from one the sum of the probabilities that w_m or w_j is correct. A similar procedure can be used for computing the correctness probability in case of consensus. The correctness probability is also a confidence measure for deciding rejection.

When the vocabulary size is small as in the case of AURORA3, interesting result can be obtained with a decision criterion based only on the prior probability $P\{C(w, w_m, w_j)\}$ as follows:

$$w^* = \underset{w \in (w_m, w_j)}{\arg\max} \; P\{C(w, w_m, w_j)\} \qquad (3)$$

5. AUTOMATIC SPEECH RECOGNITION EXPERIMENTS

ASR experiments were conducted with the test sets of the AURORA3 corpus. Only CH1 data were used for the Italian and Spanish portions. The recognizers were trained with real-life telephone data and not with the AURORA3 training corpus. Denoising was performed by non-linear spectral subtraction as described in [6].

After aligning the best word sequences generated by the two recognizers, the following decision strategy is used. If there is consensus at the word level, then the word hypothesis is validated, otherwise each word hypothesis $w_m(b,e)$ generated by the recognizer using the MRA features is considered for validation. This is motivated by the fact that this recognizer has a lower WER than the one using JRASTAPLP features.

The strategy compares $w_m(b,e)$ with the hypotheses generated by the other recognizer. Three possible cases are considered, namely:

1. *substitution:* a single word $w_{j}(b_{j}, e_{j})$ overlaps with $w_m(b,e)$ for more than 50% of the frames,
2. special case of *j-deletion*: a non-speech hypothesis is generated by the recognizer using JRASTAPLP features in the time interval of $w_m(b,e)$,
3. special case of *j-insertion:* a word hypothesis $w_j(b_j, e_j)$ is generated in a time interval where the recognizer using MRA features generates a non-speech hypothesis.

The probability $P\{C(w, w_m, w_j)\}$ was computed with the training set of Aurora 3 and used for selecting between the competing hypotheses. The decision rule is the (3). Results, in terms of WER, are reported in Table III.

Table III – Performance, in terms of WER, of the new decision strategy compared with that of the best system

	Italian	Spanish
MRA system	20.34%	15.19%
New Strategy – substitutions	6.2%	5.71%
New Strategy – deletions	9.79%	1.99%
New Strategy – insertions	1.57%	4.72%
New Strategy – overall WER	17.56%	12.42%

The WER reduction is 13.67% for Italian and 18.24% for Spanish.

6. CONCLUSIONS AND FUTURE WORK

Two ASR systems with different feature sets have been used to perform diagnosis and combination of results for improving recognition accuracy. With forced alignment of the reference sentence, it was observed that the phoneme posterior probabilities obtained with the two feature sets for the same phoneme and the same time interval may be very different in a substantial proportion of cases. This may be due to the fact that probabilities are computed by different models, by the effect of denoising or by a different discrimination power of the feature sets for different phoneme classes. Future work will investigate whether or not the trend is confirmed with different modeling techniques. If the trend will be confirmed, then attention will be paid to the features in the attempt to characterize confusions due to intrinsic variability.

A new strategy for combining ASR system results has been proposed. It is based on the probability that a hypothesis is correct, given the identity of the word hypotheses that are in competition. Significant WER reductions have been found for the CH1 proportion of the Italian and Spanish components of the AURORA3 corpus. Future work will investigate how this approach can be extended to tasks with large vocabularies.

ACKNOWLEDMENTS

The work described in this paper is part of a research effort carried out in the DIVINES project on feature extraction and diagnosis for Automatic Speech Recognition. This project is partially funded by a programme of the Human Language Technology Division of the European Community.

REFERENCES

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

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

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

[4] A. Sankar, “Bayesian model combination (baycom) for improved recognition”, *IEEE Intl. Conference on Acoustics, Speech and Signal Processing*, Philadelphia, PA, March 2005, I, pp.845-848.

[5] H. K. Kim_ and M. Rahim, “Why Speech Recognizers Make Errors? A Robustness View”, *Proc. International Conference on Spoken Language Processing,* Jeju, Korea, 2004, ThA1703o1.

[6] R. Gemello, F. Mana, D. Albesano, R. De Mori, “Multiple Resolution Analysis for Robust Automatic Speech Recognition”, *Computer Speech and Language,* (accepted, 2004).

[7] 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.