AUTOMATIC ALTERNATIVE TRANSCRIPTION GENERATION AND VOCABULARY SELECTION FOR FLEXIBLE WORD RECOGNIZERS

D. Torre, L. Villarrubia, L. Hernandez, J. M. Elvira*

Speech Technology Group, Telefónica Investigación y Desarrollo
Emilio Vargas 6, 28043 - Madrid, SPAIN
e-mail: {doro,luigi, chema}@craso.tid.es
* E. T. S. I. Telecomunicación, Universidad Politécnica Madrid, e-mail: luish@gaps.ssr.upm.es

ABSTRACT

In accordance with the new emerging Voice Response Systems that use Flexible Vocabulary Recognizers (FVRs), prediction of word confusabilities have been received increasing interest during the last few years. In this contribution we present a new method for transcription confusabilities estimation based on a new statistical modelling criterion. We propose the use of the new transcription confusability measure in two different word error rate (WER) reduction procedures for FVRs: an automatic vocabulary selection procedure suitable for those applications where the set of vocabulary words is not totally defined by the application, and an automatic procedure for generation of alternative transcriptions.

Experimental results using a telephonic database show 20% WER relative reduction using the automatic alternative transcription generation procedure for a 37 word vocabulary, and over 50% (20%) WER relative reduction using our unrestricted (restricted by groups of synonyms) vocabulary selection procedure instead of random word selection.

1 INTRODUCTON

During the last few years the idea of using recognizer's phoneme confusions information at higher recognition levels to reduce error rates has been spreading [1,2,3,4]. Phoneme confusions information has been used to estimate word confusabilities [1,2], and these estimations have been proposed to improve performance in automatic speech recognition systems. However, very few experimental results showing the error rate reduction achieved applying this idea have been reported.

In this paper we explore two of these applications: vocabulary selection and automatic alternative transcription generation.

It is well known that recognizer's error rate strongly depends on the vocabulary [1,2,3], therefore an adequate vocabulary selection results decisive for the obtained error rate. An ideal vocabulary should (a) be easy and natural for users, and (b) facilitate recognizer´s work. However, criterion (b) is difficult to apply in designing the vocabulary. This is the reason for an increasing interest in developing tools which facilitate vocabulary design on recent years. Some researchers have centred their efforts on giving an estimation of the average word error rate [1], while others try to give an estimation of word confusabilities [2]. These estimators allow to detect problematic vocabularies or words in advance of the application launching. Nevertheless, both, estimations interpretation and fixing actions to be taken continue relying on the application designer, whose experience is still crucial. In this paper we present a vocabulary design tool which requires only to have defined a group of synonyms for each 'concept' to be included in the vocabulary. Once defined the groups of synonyms our tool estimates the word confusabilities, and selects for each 'concept' the optimum synonym according to a criterion of minimum word error.

The other application we explore -automatic alternative transcription generation- uses the phonetic errors information present in the phoneme confusion matrix. These errors could be caused by either intrinsic recognizer's tendency to confuse a pair of phonemes (due to imperfect modelling), or alternative pronunciations present in the utterances used in phoneme confusion matrix training. The proposed procedure automatically includes the alternative transcriptions of a word which are:

- similar enough to the canonical transcription of that word
- different enough from the canonical transcriptions of the other vocabulary words.

2 CONFUSION PROBABILITIES ESTIMATION

Let p_i denote the phoneme labels. A phoneme transcription is a sequence of phonemes $T=(p_1, p_2, \ldots p_n)$ where n is the length of the transcription. We define a word W as a set of transcriptions, and denote that a transcription T is valid for a word W which is included in the vocabulary of the FVR as $T \in W \in VOC$.

We can obtain an estimation of $P(\text{rec } W_o \mid \text{utt } W_i) = P(W_o \mid W_i)$ from an estimation of the transcription confusion probabilities, which is in turn obtained from an estimation of phoneme confusion probabilities.

• Phoneme confusion probabilities.

We estimate the phoneme confusion probabilities matrix with an unrestricted phone recognition (UPR) experiment, using an unconstrained phone grammar and a dynamic programming matching (DP) [1]. A reestimation procedure of the DP warp function similar to the one presented in [4], but reestimating the matrix until convergence instead of reestimating the matrix once is applied. In a first step we use fixed costs for the dynamic programming (DP) matching, in successive steps we use the last estimated phone confusion matrix and a recognizer adapted DP matching to estimate the next phone confusion matrix. The recognizer adapted DP matching maximizes transcription confusion probabilities as estimated by eq.1 (discussed below). Convergence is reached in 3 or 4 steps. The problem of undertraining is avoided using a probability floor.

The obtained phone confusion matrix includes the following informations:

$P(p_o \mid p_i)$ - phoneme hit and confusion probabilities

$P(del \mid p_i)$ - phoneme deletion probabilities.

$P(p_o \mid ins)$ - phoneme insertion probabilities.

The experiment used for estimating the phone confusion probabilities implies that these probabilities are only valid for UPR experiments.

• Transcription confusion probabilities.

For phoneme insertions we assume a new statistical model:

- An insertion point is defined as the point of the uttered transcription where one or more insertions can appear. There are n+1 insertion points in a transcription with n phonemes.
- The probability of inserting once more the model p_o in an insertion point, $P_g(p_o)$, is considered statistically independent of context and previous insertions. These probabilities are obtained from the probabilities $P(p_o \mid ins)$.
- The probability of inserting once more any phone in an insertion point, P_{tg}, is the sum of all $P_g(p_o)$. The probability of inserting no more models in an insertion point is $1-P_{tg}$.

Using the confusion matrix probabilities and the above statistical model the transcription confusion probabilities between transcription T_i and T_o is estimated as:

$$P(T_o \mid T_i) = \left(\prod_{\text{all hits}} P(p_i \mid p_i)\right)\left(\prod_{\text{all subst}} P(p_o \mid p_i)\right)\left(\prod_{\text{all del}} P(del \mid p_i)\right)\left(\prod_{\text{all ins}} P_g(p_o)\right)\left((1-P_{tg})^{n_i+1}\right) \qquad \text{(EQ 1)}$$

where n_i is the number of phonemes of the transcription T_i, all hits, all subst, all del and all ins are obtained by DP matching between T_i an T_o.

The main advantage of eq.1 over other approaches comes from the fact that, apart of modelling the phoneme insertions $\left(\prod_{\text{all ins}} P_g(p_o)\right)$, we consider explicitly the probability of no (more) phoneme insertions by the term $(1-P_{tg})^{n_i+1}$.

• Word confusion matrix

Eq. 1 allows us to estimate transcription confusion probabilities in an UPR experiment, but we are interested in estimating P(Wo|Wi) in a lexical restricted recognition (LRR) experiment. There are basically two approaches to do this:

- One step algorithms, for example the one presented in [1].

$$P(Wo \mid Wi) = \sum_{\forall T} P(Wo \mid T)\, P(T \mid Wi) \qquad \text{(EQ 2)}$$

where T is an intermediate transcription between the spoken word W_i and the recognition word W_o.

- Two steps algorithms, as presented in [2]:

$$P(Wo \mid Wi) = \sum_{\forall Ti \in Wi} \sum_{\forall To \in Wo} P(To \mid Ti) P(Ti \mid Wi) \qquad \text{(EQ 3)}$$

Two step algorithms have more solid theoretical foundations, however one step algorithms give enough good estimation of word confusabilities for this task [3], saving a considerable amount of computation. Therefore, word confusion probabilities are estimated using a one step algorithm.The word confusion matrix is later normalized obtaining rows summing one.

3 ERROR REDUCTION PROCEDURES

• Vocabulary Selection (BestVoc)

This algorithm solves the following hypothetic problem: from M words we want to select the subset of N words (best_VOC) which provides the lowest word error. This is not a practical problem, but it allows us to evaluate the capacity of the proposed procedure to improve recognition rate with an adequate vocabulary choice.

The algorithm used could be stated as:

1. Estimate the MxM word confusion matrix.
2. best_score = ∞
3. for each seed_word = W_1 .. W_M
    4. tmp_VOC={seed_word}
    5. while number of words in tmp_VOC $\neq$ N
        6.obtain the word $W_k \notin$ tmp_VOC which minimizes $\sum_{\forall W_j \in tmp\_VOC} (P(W_j \mid W_k) + P(W_k \mid W_j))$
        7.tmp_VOC = tmp_VOC $\cup$ $\{W_k\}$
    8. if best_score > $\sum_{\forall W_j \in tmp\_VOC} \sum_{\substack{\forall W_k \in tmp\_VOC \\ W_k \neq W_j}} P(W_j \mid W_k)$
        9.best_VOC = tmp_VOC
        10.best_score= $\sum_{\forall W_j \in tmp\_VOC} \sum_{\substack{\forall W_k \in tmp\_VOC \\ W_k \neq W_j}} P(W_j \mid W_k)$

• **Vocabulary Selection using Synonyms (BestVocSyn)**

A practical problem in selecting vocabularies could be stated as follows: we have N groups of synonyms and we want to obtain the N vocabulary words which yields the easiest recognition task, selecting just one word of each group. If a particular word has to be present in the vocabulary, that word must be alone in its group. Otherwise, different words expressing the same 'concept' must be placed together in the same group of synonyms, allowing the algorithm to select the word which leads to lowest error.

The problem can be solved using the BestVoc algorithm with the additional restriction of selecting just one word of each group of synonyms.

• **Automatic Alternative Transcription Generation (AutoAltTrans)**

This algorithm automatically adds alternative non-canonical transcriptions taking into account the phoneme confusion matrix. The goal is to include alternative transcriptions for a word without degradation of recognition performance for the rest of the vocabulary words.

The proposed algorithm could be described as:

1. for each $W_i \in$ orig_VOC
   2. for each $T_i \in W_i$ (T_i are canonical transcriptions for W_i)
      3. obtain the set of (possibly non-canonical) transcriptions T for which $P(T|T_i) >$ THRESHOLD
      4. for each $T \neq T_i$ included in the set previously obtained
         5. if $P(T|T_i) > \text{SEC\_MARG} \cdot P(T|T_k)\ \forall T_k \in W_j$, $\forall W_j \in \text{orig\_VOC} - \{W_i\}$ include T as an alternative transcription for W_i.

The set of transcriptions T, mentioned in step 3, is calculated progressively introducing phonetic errors in transcription T_i.

The parameter THRESHOLD guarantees that the new transcription is similar enough to the canonical transcription T_i, while the parameter SEC_MARG assures that the new transcription is different enough from the transcriptions of the other vocabulary words. This control avoids the possibility of inserting transcriptions close to other vocabulary words, which would increase the error rate.

4 DATABASE

A set of 40 previously trained context-independent sex-dependent models were used. These models were trained with 5829 isolated word utterances. The phoneme confusion matrix was trained using 5078 isolated word utterances distinct from the model training ones. Results for BestVoc and BestVocSyn were obtained using 1511 utterances of 160 spanish names and surnames, for AutoAltTrans 1008 utterances of 37 spanish names and surnames were used. In both cases utterances and words are different from the ones used for training. All data used was taken from the VESTEL real telephone speech database [5].

5 RESULTS

• **Vocabulary Selection (BestVoc)**

From a vocabulary composed of 160 spanish names and surnames, subvocabularies of 10, 30 and 50 words were selected. For each subvocabulary size 24 subvocabularies were selected:

- the 6 best vocabulary candidates obtained with the BestVoc algorithm.
- 12 vocabularies composed of words randomly selected.
- the 6 worst vocabulary candidates obtained with a slight modification of the BestVoc algorithm.

The figure below represents the average error rate obtained with the 6 best (BestVoc), the 12 random (RandVoc), and the 6 worst (WorstVoc) vocabularies.

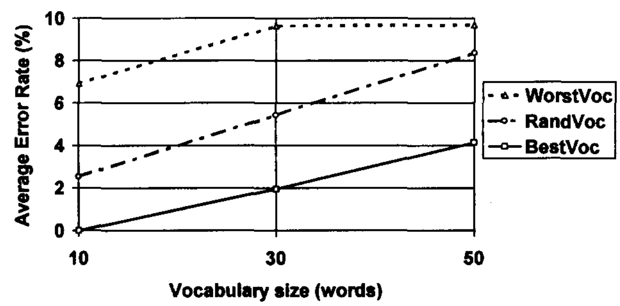

Figure 1: *Unrestricted vocabulary selection.*

This figure shows the capability of the BestVoc algorithm to select the vocabularies which yield lowest (and highest) error rates. Over a 50% average word error rate relative reduction, compared to the randomly selected vocabularies, is achieved for all vocabulary sizes, with a maximum average error rate absolute reduction of 4,21% for the 50 word vocabularies.

• **Vocabulary Selection using Synonyms (BestVocSyn)**

From the same set of 160 spanish names and surnames vocabularies of 10, 30 and 50 words were formed. The process used to select these vocabularies is represented in figure 2, where N is the vocabulary size.

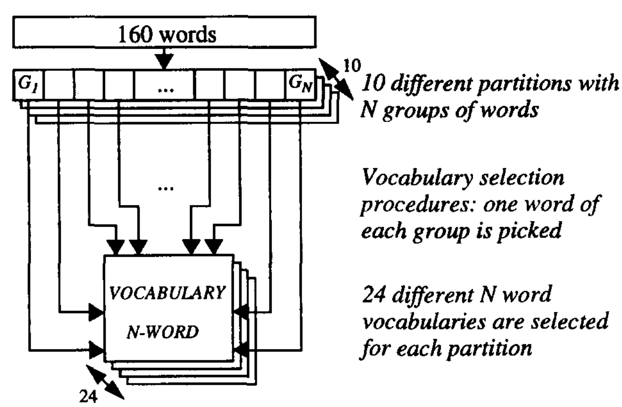

Figure 2: *Vocabulary selection test using N groups of words.*

First, 10 different partitions obtained, each one, by splitting the set of 160 spanish names and surnames in N groups of equal number of words were formed. Therefore, each partition contains N groups of equal number of words which simulate the synonym groups in a real application. The restrictions imposed by the groups of words limit the capacity of the selection algorithm to reach both, the best and the worst vocabulary, but this effect depends on the selected groups of words. Using 10 different, randomly generated partitions the effect of the particular choice of groups is reduced.

For each of these partitions 24 vocabularies were selected by picking one word of each group:

- the 6 best vocabulary candidates given by the BestVocSyn algorithm
- 12 vocabularies composed with one randomly selected word of each group
- the 6 worst vocabulary candidates given by a slight modification of the BestVocSyn algorithm.

This process was repeated with N = 10, 30 and 50 word vocabularies. For each vocabulary size three points were represented in figure 3: the average WER obtained in recognition tests using the 60 best (6 best candidates given by BestVocSyn for each of the 10 partitions used), the 120 random (RandVocSyn), and the 60 worst (WorstVocSyn) vocabularies.

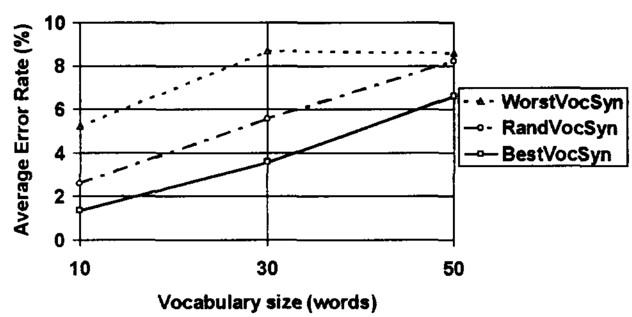

Figure 3: Vocabulary selection using groups of synonyms.

The restrictions imposed by the groups of synonyms limit the capacity of the selection algorithm to reach both, the best and the worst vocabulary, thus reducing the margin between error rates obtained with the best and the worst vocabularies. Nevertheless, the figure shows that, even with these restrictions the error rate margins are still important: the BestVocSyn algorithm yields over 20% average WER relative reduction for all vocabulary sizes, compared to random vocabulary selection, with a maximum average WER absolute reduction of 2% for the 30 word vocabularies.

• Automatic Alternative Transcription Generation (AutoAltTrans)

We have make a preliminary test with a vocabulary of 37 words, obtaining the error rates indicated in Table 1 for the three tested cases:

(a) grammar with just canonical transcriptions (sex-dependent).
(b) grammar with canonical and non-canonical automatically added transcriptions, summing about twice the number of transcriptions in (a)
(c) grammar with canonical and non-canonical automatically added transcriptions, totalling about eight times the number of transcriptions in (a)

Test Case	THRESHOLD	SEC_MARG	Number of Transcriptions	Error Rate (%)
a	---	---	74	6,45
b	2,5	31,6	177	5,56
		3,16	179	5,56
c	1,17	31,6	630	5,26
		3,16	653	5,16

***Table 1:** Automatic alternative transcriptions generation.*

Table 1 shows an average word error rate relative reduction of 20% for the maximum number of included transcriptions. These results show that it is possible to improve recognition rate by means of including alternative transcriptions automatically generated having into account the phoneme confusion matrix and the new proposed method for estimate transcription confusability.

6 CONCLUSIONS AND FUTURE WORK

A new procedure for transcription confusabilities prediction has been presented, and two different error rate reduction procedures have been tested: an automatic vocabulary selection procedure and an automatic procedure for generation of alternative transcriptions. Experimental testing has shown that a 20% word error rate relative reduction can be achieved with these procedures. However, further testing with larger vocabularies would be desirable. Finally, we are working in combining automatic and knowledge based generation of alternative transcriptions in order to increase the recognizer performance.

REFERENCES

[1] Alison Simons. "Predictive Assessment for Speaker Independent Isolated Word Recognisers", Proc. EUROSPEECH 95, pp. 1465-1567.

[2] David B. Roe, Michael D. Riley. "Prediction of Word Confusabilities for Speech Recognition", Proc. ICSLP 94, pp. 227-230.

[3] Borge Lindberg. "Recogniser Response Modelling from Testing on Series of Minimal Word Pairs", Proc. ICSLP 94, pp. 1275-1278.

[4] C. Bourjot, A. Boyer, D. Fohr. "Phonetic Decoder Assessment", Proc. EUROSPEECH 89, pp. 457-460.

[5] D. Tapias, A. Acero, J. Esteve, J.C. Torrecilla. "The VESTEL Telephone Speech Database", Proc. ICLSP 94, Yokohama, pp. 1811-1814.