CREATING USER DEFINED NEW VOCABULARIES FOR VOICE DIALING

José María Elvira, Juan Carlos Torrecilla, Javier Caminero.
Speech Technology Group
Telefónica Investigación y Desarrollo, Emilio Vargas 6, 28043 Madrid, Spain
e-mail: (chema, jcarlos, jcam)@craso.tid.es

ABSTRACT

This paper introduces a new approach for generation of phonetic transcriptions for voice dialing applications. where on-line construction of user vocabularies is mandatory. The proposed method allows adaptive selection of new transcriptions requiring much less speech utterances for system training than other approaches. The new approach is compared to other classical approaches showing a clear improvement on performance and efficiency.

1. INTRODUCTION

All speech recognition systems based on subword or phone-like units need of a preparation step where the vocabulary words to be recognised are transformed into the subwords or phone-like strings used in the recognition. The collaboration of a phonetician or a specially prepared software is needed to undertake this transformation. Nowadays, most of the speech recognition applications use flexible or voc-independent speech recognisers that can not be modified on real-time. If any modification is required, the system has to be stopped, the modification done and the system restarted again. However, there exist applications where the modifications are frequent and needed and the system can not be stopped. One application of this kind is a personal telephony directory for voice dialing. Adding, erasing or reviewing names will be a usual process. Of course, it would not be practical to have a system manager to do all these operations for all the final users. It would be much better allow the user to do all this by himself using a simple telephone interface.

A solution to this problem is the development of a system able to take speech examples of the words to be added and insert the transcriptions of these examples into the vocabulary of the recogniser. Some strategies have been developed in the literature for this propose [3][4], however, the problem arises when the most suitable transcriptions have to be selected. These classical methods show some deficiencies in the results obtained; increment of the WER (Word Error Rate) when new speech examples are used, or when more transcriptions are added is frequent.

This work introduces a new approach for new word addition in dynamic vocabularies. This approach uses two phases: a transcriptions generation process and a transcriptions selection step. A feedback parallel grammar with different sub-word models and a contextual bigram is used for the generation process. The transcriptions selection step has been designed to avoid interferences between already existing transcriptions, and to use just the necessary speech examples to obtain the most suitable transcriptions. This selection process is based on the use of a new distance measurement between transcriptions.

The structure of this work is as follows. Initially, the speech database, the HMM models and the strategy for transcription generation is presented. After this, the classical force-alignment method is presented together with some results. Finally, the proposed method is evaluated and compared to the forced alignment method, to finish with some conclusions.

2. DATABASE AND MODELS

For this work, data from the VESTEL [1] database was used. A set of 11 common Spanish firstnames (Ana, Juan, Luis,...) was chosen as the new vocabulary. For training, 66 utterances from different speakers (6 per new word) were taken. The evaluation data set consists of 1125 examples of the same words from different speakers. Although the system will be, a priory, speaker dependent oriented, the evaluation undertaken for these experiments is speaker independent, being a much harder task.

Two different sets of Speaker Independent CHMM (Continuous Hidden Markov Models) were used in these experiments. A first set of CI (Context Independent) models, and a second set compound of left side biphone models [2].

3. WORD TRANSCRIPTIONS GENERATION AND RECOGNITION

The transcriptions were generated using a parallel grammar with feed-back of the subword CHMM models (Figure 1). Therefore, the sequence obtained is the one that models the speech signal with a higher probability. The method is based on the uniformity of the sequences rather than their exactitude. What is important is the similarity between the transcriptions obtained from the different speech examples of the same word. Thus, the system will work with transcriptions that are not exact phonetically.

The transcription generation process is improved, as the results will present, by the addition of contextual bigrams and heuristically built bigrams into the transcription generation grammar.

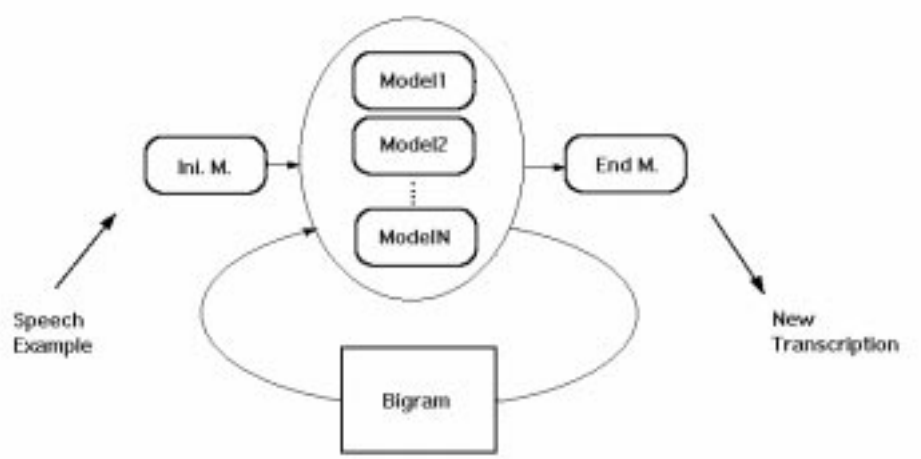

Figure 1. Transcription Generation.

# trans. per word	# of training speech examples											
	1		2		3		4		5		6	
	CI	Biph	CI	Biph	CI	Biph	CI	Biph	CI	Biph	CI	Biph
1	16.09	8.71	15.47	3.82	16.44	4.18	14.67	4.00	7.29	3.73	7.20	3.47
2			13.42	3.91	14.13	4.36	8.98	3.47	7.91	3.11	7.11	2.93
3					11.91	3.64	10.58	3.47	5.69	2.76	7.11	3.02
4							9.87	3.64	6.40	2.93	5.69	3.11
5									7.20	3.02	6.04	2.76
6											6.76	3.11

Table 1: WER Results for the forced alignment

The generated transcriptions are evaluated for the particular selection approach, and the selected transcriptions added to the recognition grammar of the new vocabulary.
The same Speaker-Independent models, along with the grammar obtained during enrolment are used for recognition.

4. FORCED ALIGNMENT; EXPERIMENTS AND RESULTS.

This method is a classical approach [3] [4]. The idea behind is to force-align each transcription (generated with the above described technique) with all the speech examples used for training that new word. The transcription with the maximum average Viterbi score is selected as the new word-model.
This approach can be extended and it can use more than one transcription. Instead of using only one transcription per new word, the system can allow the selection of more transcriptions for the new word. Therefore, depending on the number of speech examples used for training, the system can select 1 transcription, or 2, or 3,... up to the number of speech examples.
Table 1 shows the WER results obtained with this method for CI models and Biph. (Biphone) models. These results show the expected performance improvement by the use of the contextual models.

4.1. Including bigram grammar

An improvement of this method is the inclusion of a bigram grammar in the transcription generation process. For this experiment, a heuristically built bigram grammar was used. For the CI models this bigram grammar was obtained from transcribed text. For the biphone models the bigram only allows transitions between consecutive context units, that is, for a particular biphone, let say *-a (any biphone formed with the /a/ and any context) only can be follow by a-* (any biphone with left context /a/). The new results are presented in Table 2.
As it can be seen, the inclusion of the bigram grammar provides an important improvement.

# trans. per word	# of training speech examples											
	1		2		3		4		5		6	
	CI	Biph	CI	Biph	CI	Biph	CI	Biph	CI	Biph	CI	Biph
1	14.22	8.36	12.71	2.84	15.29	3.02	14.67	3.02	7.56	2.58	7.82	2.58
2			9.96	3.11	12.27	2.67	8.80	2.58	5.69	2.40	7.56	2.13
3					8.71	2.76	8.36	2.58	4.98	2.13	6.22	1.96
4							7.73	2.58	5.33	2.22	4.80	2.04
5									4.80	2.22	5.16	2.22
6											4.80	2.31

Table 2: WER Results for the forced alignment using heuristic bigram for the CI models and contextual bigram for the biphone models.

5. NEW DISTANCE BASED APPROACH

From the results presented in the two tables above it is clear that the biphone models perform much better than the CI models. Also, it is clear the improvement when the bigram grammar is introduced. However, there are other points that are not so clear. It is difficult to define the number of transcriptions required per each new word. Although there is an improvement when this number is increased, there is a point where the improvement is clearly small. On other side, the use of more than 3 transcriptions looks excessive. It has to be considered that for this kind of applications, the speech examples used for training have to be introduced by the user, then, more than three examples will be incommodious.

A different point is the convenience of using the same number of speech examples per each new word. Some new words could have enough just by using 2 speech examples, while others can need more examples for fine tuning. In some of the results, it can be seen how after new speech examples are introduced the system performance is still the same or even worse due to the variability introduced with the new examples. This can be due to the distortion of these new transcriptions between the same word examples or with the other vocabulary words. Taking all these points in consideration, a new approach based on the EAPTE method [5] was developed.

The EAPTE method offers the possibility of representing the confusion probability between two transcriptions of the same word (or different words). This transcription confusion probability is in turn obtained from an estimation of phoneme confusion probabilities. It is a simple step to transform the context dependent transcription obtained using the transcription generation process described above into the simplest phoneme transcription.

The inverse of this confusion probability can be considered as a distance factor. Therefore, by means of using this distance, an adaptive selection algorithm can be defined. This algorithm, allows the selection of the most similar transcriptions as the most appropriate transcriptions for a new word. Also, the selection process can stop when two (or N) similar enough transcriptions had been obtained.

An example of this algorithm to introduce a new word into the vocabulary using two transcriptions can be:

```
valid=FALSE
While NOT valid do
        obtain new transcription
        If new trans. similar to an old one
                valid=TRUE
end
insert the new two transcriptions.
```

This new approach was evaluated to obtain two transcriptions per new word and the results obtained are presented in Table 3 in comparison with the results obtained for the forced alignment in the same case (selection of two transcriptions from N training speech examples).

This table shows that this approach obtains similar results (when no better) using a smaller number of training utterances than the forced alignment method.

The performance was even improved adding the heuristically obtained bigram grammar into the biphone models network for the transcription generation process. This bigram improved the transcription generation and the final recognition.

Results of this new experiment are presented in Table 4, comparing, again, to the forced alignment with the same conditions.

N, max. # trans. per word	Proposed Method		Forced Alignment	
	WER %	average training utterances	WER %	average training utterances
2	3.11	2	3.11	2
3	2.67	2.6	2.67	3
4	2.67	2.8	2.58	4
5	2.22	3	2.40	5
6	2.22	3.1	2.13	6

Table 3: Comparison between the distance based approach and the forced alignment.

N, max. # trans. per word	Proposed Method		Forced Alignment	
	WER %	average training utterances	WER %	average training utterances
2	3.20	2	3.20	2
3	2.22	2.6	2.40	3
4	1.87	2.9	1.78	4
5	1.96	3.1	1.96	5
6	1.87	3.3	2.13	6

Table 4: Comparison between the distance based approach and the forced alignment using biphones and the heuristic bigram grammar.

This technique will ask for new examples only with those words where there is not a similarity between examples. Therefore, on average, the number of examples required will be much smaller.
Also, and on a practical implementation, the algorithm can be improved introducing a global evaluation. That is, the new transcriptions selected can be compared to the other vocabulary words and, in case they are too close to anyone, advise to the user for a change in the word to be introduced. Therefore, the above algorithm can be rewritten as:

```
collision=TRUE
While collision do
        valid=FALSE
        While NOT valid do
                obtain new transcription
                If new trans. similar to an old one
                        valid=TRUE
        end
        If transcriptions similar to others then
                advise change
        else
                insert the new two transcriptions.
                collision=FALSE
        end
end
```

This final algorithm has been used in all the experiments presented in this work.

6. CONCLUSIONS

This work presents and evaluates a new approach for new word addition in dynamic vocabularies. The results presented show a very good system performance in a speaker independent task. The system performs very similar to the classic approaches but it requires much fewer speech examples (see Table 3 and Table 4), increasing the acceptability of the system. The fine tuning of the parameters that control this new approach can allow to outperform the classical methods.

7. REFERENCES

[1]. D. Tapias, A. Acero, J. Estevez y J.C. Torrecilla, “The VESTEL Telephone Speech Database”, ICSLP-94, Japan, p. 1811-1814.

[2]. L. Villarrubia, L.H. Gómez, J.M. Elvira, J.C. Torrecilla, “Context-Dependent Units for Vocabulary-Independent Spanish Speech Recognition”, ICASSP’96, p. 451-454.

[3]. R. Haeb-Umbach, P. Beyerlein, E. Thelen, “Automatic Transcription of Unknown Words in a Speech Recognition System”, ICASSP’95, Detroit 1995, p. 840-843.

[4]. J. Neena, R. Cole, E. Barnard, “Creating Speaker-Specific Phonetic Templates with a Speaker-Independent Phonetic Recognizer: Implications for Voice Dialing”, ICASSP’96, p. 881-884.

[5]. D. Torre, L. Villarrubia, L. Hernandez-Gómez, J. Elvira, “Automatic Alternative Transcription Generation and Vocabulary Selection for Flexible Word Recognizers”, ICASSP’97, p. 1463-1466.