NAME DIALING USING FINAL USER DEFINED VOCABULARIES IN MOBILE (GSM & TACS) AND FIXED TELEPHONE NETWORKS

J.M. Elvira, J.C. Torrecilla

Telefónica I+D

Speech Technology Group
Emilio Vargas 6, 28043 Madrid Spain.
e-mails: (chema,jcarlos)@craso.tid.es

ABSTRACT

This paper presents the results obtained on the evaluation of a new approach for generation of phonetic transcriptions for name dialing applications in different telephone networks and with temporal variations. In this kind of applications 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 evaluated using data from different telephone networks (PSTN, GSM and TACS networks) and from different temporal utterances (recordings done in a period of two months).

1. INTRODUCTION

All speech recognition systems based on subword or phone-like units need of a preparation step where the vocabulary words to be recognized 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 vocabulary 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 name 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.

A recent work [6] 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 [5] and, as presented in [6] is developed as the following algorithm to add a new name in the directory or vocabulary:

```
collisions=TRUE
While collisions 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
```

where a collision is produced when a similar transcription is already in the vocabulary.

Other problem added to these kind of applications is the difficulty on knowing where the user is calling from or the kind of telephone network where the call is coming through. Therefore, the system has to be able of working properly in mixed network environments.

At the same time, and due to the methodology used, another factor that has to be evaluated is the robustness of the system with the temporal variation of the user speech.

This work evaluates the early proposed method [6] on applications of this kind.

The structure of this paper is as follows. Initially, the speech database and the HMM models are introduced. Next a description of the experiments reported in this work is given. Later, the results obtained in the experiments are presented, to finish the paper with some conclusions.

2. DATABASE AND MODELS

For the experiments reported in this work a especial database was assembled. The corpus of the database consists of a set of 100 different words or short phrases. 50 of the words were a selection of the most frequent names found in the VESTEL [1] database. 30 were short phrases build from the concatenation of the 10 most frequent words from VESTEL and 3 different suffixes (“movil” (mobile), “casa” (home), “oficina” (office)), and the other 20 were especial words than were considered as very common in a personal agenda. Words such as “doctor”, “ambulance”, were part of this last subset. 30 different vocabularies using 25 different words each from the corpus were assembled. Therefore, the apparition of similar words in the vocabularies was very common.
30 speakers were used to record speech examples from each vocabulary. The recordings were done on two different sessions in an interval of 2 months. In the first session, 10 different utterances for each word were recorded. In the second, 5 utterances from each word were recorded. From the first session recordings, 5 utterances were used as training utterances and the other 5 as test utterances. Therefore, the database has three different subsets: a training set and two test sets. Also, the two evaluation sub-databases allow the evaluation of the temporal variation in the users speech.

Afterwards, the recordings were put through three real telephone networks and recorded again. Three new databases were obtained after this process. The telephone networks used were the Public Switched Telephone Network (PSTN), and the two mobile networks, the digital network (GSM) and the analog mobile network (TACS). This methodology allows the evaluation of the system behavior on applications where a possible mixed network environment can be found.

For system training (users directory assembling) and recognition, the system uses a set of context dependent models presented in other works [2][6]. The models are a set of CHMM (Continuous Hidden Markov Models) compound of left side biphone models. These models are speaker independent models trained using the VESTEL database [1] that is a database recorded using the PSTN. This is also another factor to evaluate, the behavior of a set of models trained in a particular telephone network, working on different telephone networks.

3. EXPERIMENTS

Three different experiments were undertaken.

- The first step was the construction of the recognition vocabularies, or user directories, using the training subset from each telephone network. Therefore, user directories were assembled for each telephone network
- The second experiment was the evaluation of each user directory using the corresponding test examples from both sessions.
- The third and last was the evaluation of the user directories crossing the telephone network databases.

All these experiments allow the evaluation of the different aspects of the system. The first experiment gives results about the system training process. It also informs about the system behavior through the different telephone networks, even the models were obtained from the PSTN network.

The second experiment gives information about the general system behavior taking into account the variability of the user voice through a moderate period of time.

And the third experiment introduces results about the system behavior when it is used on a cross network application together with the variability of the user voice through a moderate period of time.

4. RESULTS

In the first experiment (system training) the algorithm used was the one presented on [6] as the practical

algorithm. The consideration that has to be taken in account is that when a collision was detected the system is allowed to use the transcription. Then, after this modification the algorithm is as follows:

```
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
        ncollisions++
insert the new two transcriptions.
```

Using this strategy the results obtained during the training or directories generation was as presented in Table 1 for the different telephone network.

Network	speech examples used			average examples per word			collisions		
	m	a	M	m	a	M	m	a	M
PSTN	56	63.8	73	2.24	2.55	2.92	1	3.5	8
GSM	56	65.3	80	2.24	2.61	3.20	1	3.4	7
TACS	54	61.2	73	2.16	2.45	2.92	1	3.7	7

Table 1. Training results for the different networks.

Table 1 shows the values obtained where the tree mayor columns stands for the most important factors during training. The first column indicates the number of speech examples required for the system to build each user directory (where “m” stands for the minimum value, “a” for the average value and “M” for the maximum value). The second column indicates the average number of examples required per word (with the minimum average and maximum values) and the third column stands for number of collisions.

From this table a conclusion can be obtained. The training behavior is very similar for the three telephone networks. the best behavior is observed for the Analog Mobile Network were the average number of training utterances required is small although the differences are very small. However, there is a factor that has to be considered, due to the methodology used to build the databases there are no fading effects in the mobile databases.

The results obtained for the second evaluation experiment are presented in Table 2. This experiment, as mentioned before, evaluates the system behavior for the matched telephone networks, that is, the examples used to evaluate the system in the name dialing process are examples recorded through the same network than the one used to build the users directories. Taking in account the evolution introduced in the speech examples due to the temporal separation of the two test sessions.

Network	first session	second session	average
PSTN	1.21	1.80	1.50
GSM	1.10	2.36	1.73
TACS	0.38	1.27	0.82

Table 2. Recognition WER% using matched network test databases.

This table shows clearly the increment in the WER from the first session to the second. Two reason were found for this change. The speakers, in the second session, were more familiar with the recording procedure. Due to this, they put less care. The other reason was the change in the speakers voices after two months. This two problems are difficult to fix in the results obtained. There exist speakers were the WER increment is negative or almost almost zero but there exist other where the increment is very important. Table 3 shows two of these examples for the PSTN network.

speaker	first session	second session
fj	0.82	0.00
fb	0.00	1.60

Table 3. WER for two speakers

Table 3 show the two different behavior, although the increment in the WER is the most general.

However, the system behavior was almost perfect. The worst WER factor was 2.36% for the Digital Mobile Network (GSM) in the second session (Table 2).

The last results show the system behavior when the telephone networks were crossed, that is, once the system is trained using a network, the system is tested using data from the other networks.

These results are presented in Table 4.

training network	test network	first session	second session	average
PSTN	GSM	2.21	2.50	2.35
	TACS	2.19	2.43	2.31
GSM	PSTN	1.67	2.65	2.16
	TACS	0.99	2.34	1.66
TACS	PSTN	1.45	2.49	1.97
	GSM	0.97	1.64	1.30

Table 4. Recognition WER% using crossed network databases for test.

If Table 4 and Table 2 are compared, there is a clear WER increment on average. That was something expected. However, the average behavior is more than acceptable. The worst WER was 2.65% (Table 5). These results show that the most robust network is the Analog Mobile Network (TACS) that presents the best results in both tests.

5. CONCLUSIONS AND FUTURE WORK

This work evaluates a new approach for new word addition in dynamic vocabularies using real data from different telephone networks for name dialing applications. It also evaluates the performance variation due to the temporal evolution of the user voice. The results presented show a very good system performance.

The system performs very similar for the three different telephone networks although the models used were designed using the PSTN network. The TACS network shows the most robust behavior.

The other factor evaluated, the effect of the time evolution of the voice shows a lees clear behavior. The WER increment looks to be the general evolution but some speakers shows a different conduct.

Two different points can be evaluated on future experiments. The introduction of fading effects in the recording methodology will help to create more real databases. On the other side, the inclusion of a third session in the evaluation process with a longer temporal evolution can help to clarify the speaker voice evolution effect.

6 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.

[6] J.M. Elvira, J.C.Torrecilla, J. Caminero, “Creating User Defined New Vocabularies for Voice Dialing”, EUROSPEECH’97, p. 2463-2466.