

Introducing a Framework to Create Telephony Speech Databases from Direct Ones

Saeedeh Momtazi¹, Hossein Sameti², Saman Vaisipour³, Meysam Tefagh⁴

Department of Computer Engineering

Sharif University of Technology

Azadi st., Tehran, Iran

E-mail: ^{1,3,4}momtazi, vaisipour, tefagh@ce.sharif.edu

²sameti@sharif.edu

Keywords: Speech Recognition, Telephony Speech Recognition, FARSDAT, Speech Databases.

Abstract - A Comprehensive speech database is one of the important tools for developing speech recognition systems; these tools are necessary for telephony recognition, too. Although adequate databases for direct speech recognizers exist, there is not an appropriate database for telephony speech recognizers. Most methods suggested for solving this problem are based on building new databases which tends to consume much time and many resources; or they used a filter which simulates circuit switch behavior to transform direct databases to telephony ones, in this case resulted databases have many differences with real telephony databases. In this paper we introduce a framework for creating telephony speech database from direct ones in order to reduce the costs of other existing methods. We apply this framework to FARSDAT and produce a telephony database which was used in a telephony command recognizer.

1. INTRODUCTION

One of the most important required tools for speech recognition is a database which is used for training the acoustic models. Since speech recognition systems are intended to be speaker independent, these databases must contain utterances which are uttered by various speakers with different ages, genders, accents and educational level. Based on these assumptions FARSDAT was built in 1996 by "Research Center of Intelligence Signal Processing". FARSDAT is the first Persian language database that is built for examining and modeling acoustic features of Persian language phonemes. This database includes 386 sentences that contain 1200 most frequent words of Persian language.

All sentences in this database are expressed by 304 different people with different ages, genders, education levels and accents. All speakers who were selected to express FARSDAT sentences had one of 10 common Persian accents. Every speaker pronounced 20 sentences which among them 18 sentences were various for each speaker and 2 remaining sentences were common for all speakers. All of the sounds are recorded in an acoustic room by a 16 bit sound blaster card. The sampling rate is 44.1 kHz and signal to noise ratio is about 32 db. [1]

Although this database is adequate for direct speech recognizers, it is not appropriate for telephony speech recognizers which deal with so different speech signals. There are 2 solutions for solving this problem:

- Building a new telephony database

- Telephonizing direct databases

Both solutions are under development now; the first telephony database naming Telephony FARSDAT was made in 1999. This database contains informal and formal speeches. Despite of FARSDAT, there is not enough variety of speakers in this telephony version. Also the formal speeches that are expressed in this database are not in the continuous form and are only some numbers, days and months name and Persian alphabet.[2] The second telephony database naming Large Telephony FARSDAT is under development now and will be ready soon.[2] Although this telephony database is more complete than the first one, it is not as complete as FARSDAT. Also all signals of this database are recorded by a special device which may caused some degradation in recognition accuracy if we use other devices during the test phase. Beside this, the procedure of building a new speech database is difficult and time consuming. For example building the Large Telephony FARSDAT was started in 2003 and it is not finished until now.

According to these problems, better alternative is using existed direct databases for building the telephony ones. There exist two methods for telephonizing current direct databases; using an artificial filter which simulates the telephone networks and transferring of all signals through a real phone line. In this paper a comprehensive structure for implementing second method is prepared and a telephony database was produced by transferring FARSDAT through this structure. We compare our works with another database produced from affecting FARSDAT by an artificial filter.

The structure of this paper is as follows. Our framework for converting direct speech databases to telephony ones is discussed in section 2. Section 3 includes the implementation of this new method and its specification for FARSDAT by using dialogic board and external modem. The experimental results in continuous speech recognition system are discussed in section 4. Finally, in section 5, concluding remarks are made.

2. SYSTEM CONFIGURATION

We set up a configuration to telephonize a speech database using some kinds of hardware devices; our proposed method can be changed easily to achieve better performance over any specific hardware. One of the challenging issues in this task is its precision, because speech databases should be transferred through the circuit switch network so elaborately that its annotated

transcription would not be changed. If the transferring procedure does not preserve this property then the resulting speech signals are not useful because they do not have any correct annotated transcription.

The transferring procedure of the speech database was done via 5 distinct steps which are shown in Figure 1. First, we reduced the bit rate of speech signals by both down sampling and reducing modulation size; these changes are done according to the specification of the device which is used for playing speech signals. Also, according to the bandwidth limitation of phone lines which is about 4 kHz, more than 8 kHz sampling rate is not appropriate for transferring. After wave files are changed to a desired format, we concatenate them to produce bigger wave files which are more suitable for transferring through the circuit switch networks. This step is necessary for databases which are composed of many short signals. Length of concatenated files may change from few minutes to several hours based on the size of the database and hardware limitations. Note that the concatenation process of wave files should be logged for the future splitting step. Now these wave files are ready to be played by a telephony device after we add some specific signals to the both side of speech signals; these additional short signals are necessary for two reasons. The first one is protecting speech signals from losing data in first or last points and the second one is for facilitating of finding exact start and end points of recorded speech signals. We add a sinusoid signal with a constant frequency to the start and end points of speech signals, by this mean start and end points of recorded signals can be founded by counting some picks at each side.

Final transferred speech signals can be retrieved by performing above steps on the recorded wave files, reversely. At the first, additional sinusoid signals should be removed from the start and end points of files and then they should be split based on the logs which are saved during the concatenation step, by the way final transferred speech signals are produced.

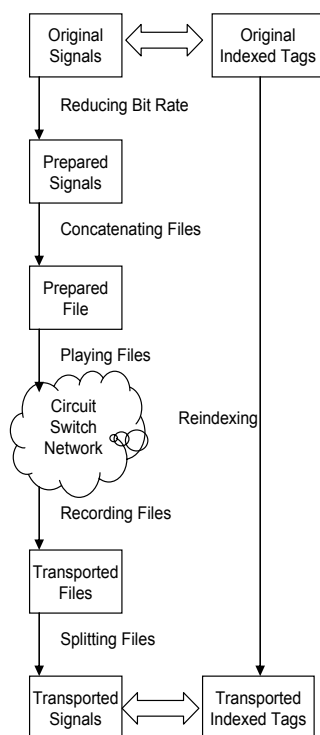


Fig. 1. Our Framework

As above steps are performed for transferring speech signals, a simple action is needed for preparing transcription associated with transferred signals. Only action which may be needed for transcriptions is a "Reducing Bit Rate" procedure. If a down sampling action is done in the first step and also annotations are based on samples number then a "re-indexing" step is needed to prepare it for new down sampled speech signals.

3. IMPLEMENTATION

As we said, FARSDAT contains utterances from 304 different speakers with various ages, genders, education levels and accents that each of them expressed 20 sentences. Each speaker's sentence is saved as two files. So there are 608 wave files which consist of 9-12 short Persian sentences. [1]

For saving analyzed data the hierarchical technique is used. For this reason beside each wave file there is a text file that terminates by ".snt". The scope of each sentence is saved in text file and each sentence has an index that is started by 1.

For each sentence in wave file, there is another file named like the wave file and finished by the sentence number. In this file the words of each sentence are analyzed according to the sample's number. Also there is another analysis for each word and they are broken into several phonemes which are saved in another file. This hierarchical structure helps us for searching different items in this set, but it is not suitable for phoneme analysis of a file because it needs some indirect levels.

For this reason we change this structure to a simple one which is suitable for our speech recognition engine. In this structure each of 6080 sentences are saved in a separate wave file and there is a .phn file beside each wave file which contains the phoneme transcription of its corresponding utterance.

To facilitate transfer of wave files from circuit switch network we divide them into 13 groups and concatenate files of each group together. So instead of 6080 files with approximately 2 second length, 13 files with 20 minutes length should be transferred.

Another important change which we used in our project is changing the resolution and sampling rate of each wave file. The sampling rate of FARSDAT is 22050 Hz and each sample is represented by a 16 bit digit. For adapting these files to phone lines we should down sample them to 8000 Hz with 8 bit resolution.

After above steps our wave files are ready for transferring from phone lines. For this task we use "Intel Dialogic Board" and "External Voice Modem" with voice capability for transition. Device Specifications listed in the following table [3].

Table1. Device Specifications

Device	Dialogic	External Modem
Vendor	Intel	D-Link
Device ID	D/4 PCIU	DFM-560EL
Play	Sample Rate	8000
	Modulation	16-bit
	Coding	PCM
Record	Sample Rate	8000
	Modulation	8-bit
	Coding	PCM

“Dialogic Programming API” is used to drive Dialogic Board playing and recording waves. [4] Also host programming language on dialogic sides is C# .NET.

We use AT Commands described in “TIA/EIA 602 standard” [5] to communicate and drive the Modem and “IS-101 AT+V commands with extensions” [6] communicating modem in voice activated mode. Also host programming language on Modem sides is JAVA 1.4.

Properties of telephony lines used for implementation listed below. Note that these parameters measured by performing a modem data handshake over lines and querying modem for line quality with special AT command. Although most of these measures vary from a call to another but may introduce some overview of line quality.

Table2. Dialogic to Dialogic Line Properties

Receive Signal Power Level	(-dBm) 25
Transmit Signal Power Level	(-dBm) 10
Round Trip Delay	(msec) 13
Near Echo Level	(-dBm) 16
Far Echo Level	(-dBm) 51

Table3. Dialogic to Modem Line Properties

Receive Signal Power Level	(-dBm) 24
Transmit Signal Power Level	(-dBm) 10
Round Trip Delay	(msec) 13
Near Echo Level	(-dBm) 16
Far Echo Level	(-dBm) 50

As shown in table 1 the playable format in dialogic board defer from FARSDAT original format, so we used “cool edit” software to reduce sample rate and modulation precision. This process needs re-indexing phonemes and we did that by applying a re-indexing procedure on all .phn files.

4. RESULTS

We tested our system from two different perspectives, in first test series we examine phoneme recognition accuracy of databases and in the second test series we examine databases in a phoneme-based isolated word recognizer (IWR). By performing first test we were assured that transferring procedure did not affect database and disorder it such that original transcriptions can not be used any more. Second test assured us that telephonized database and real telephony samples are similar enough and our transferring procedure produced a database which is

similar to a real telephony speech database. In all tests we compare four different speech databases which specification of them is depicted in the following table.

Table4. The format of different related databases

Speech Database	Sampling Rate (Hz)	Modulation Resolution (bit)
Original FARSDAT	22050	16
Reduced FARSDAT	8000	8
Transferred FARSDAT (dialogic to dialogic)	8000	8
Transferred FARSDAT (dialogic to modem)	8000	8
Tel FARSDAT	8000	8

Note that Reduced FARSDAT is produced from original FARSDAT by reducing its sampling rate and modulation resolution, this database is produced only for fairer comparison of tests. The Tel FARSDAT is the telephony database created by transferring the original FARSDAT from a special filter by “Research Center of Intelligence Signal Processing”. In the following sections our test conditions are described in details.

4.1 Phoneme Recognition Accuracy

As was mentioned, main contribution of this test is to assure that correspondence between utterances and annotated transcriptions did not lose during the transfer procedure. To achieve this goal we splitted each database into two parts, train set and test set. We used train samples to train a recognizer and then made a recognizer to recognize test utterances, we expect transferred database approximately results as good as original database.

Our recognizer is phoneme-based, i.e. there exist a distinct model in system for each phoneme of language. Models which are used in our system are Hidden Markov Models (HMM) [7] with 6 states and 8 Gaussian mixtures in each state. There are 30 HMMs in system, as a model for each phoneme of Persian language. During the training phase parameters of models are adjusted using their related train utterances.

After models were trained, test phase was started and test utterances were fed into the recognizer. Result of recognition task was a string of phonemes. We evaluate this result by comparing it with true phoneme transcription of utterances. There may be three different kinds of inequality between recognized string and reference string which are known as Insertion (I), Deletion (D) and Substitution (S). Based on these values two metrics are defined for evaluating performance of recognizers. Accuracy metric is difference between reference string length and sum of Insertion, Deletion and Substitution errors normalized by reference string length. The other metric named Correctness is like accuracy except that Insertion error is ignored in this metric. Following relations show definition of these metrics

$$Accuracy = \frac{L - (D + I + S)}{L} \quad (1)$$

$$Correctness = \frac{L - (D + S)}{L} \quad (2)$$

By using these metrics we can evaluate speech databases and consistency with their transcriptions. Following table shows Accuracy and correctness values for all 5 databases of table 4.

Table5. Accuracy and Correctness of speech databases

Speech Database	Accuracy	Correctness
Original FARSDAT	69.82	77.21
Reduced FARSDAT	59.44	67.13
Transferred FARSDAT (dialogic to dialogic)	59.67	65.36
Transferred FARSDAT (dialogic to modem)	48.87	57.46
Tel FARSDAT	23.38	38.58

As it can be seen there is a gap between accuracy of Original FARSDAT and Reduced FARSDAT, this is a common consequence of bit rate reduction because some useful information is lost during the reduction process. Another notable point in this table is negligible difference between accuracy of Reduced FARSDAT and accuracy of Transferred FARSDAT (dial to dial), this is a good evidence that is showing there was not any problem during the transferring process. The database which is recorded by modem performed worse than dialogic recorded database, this result is justifiable when we compare files which are recorded by dialogic and modem.

4.2 Word Recognition Accuracy

In this series of test we try to find how much our telephony speech databases are similar to real telephony samples. To find this we gather some real telephony commands and made a phoneme-based isolated word recognizer (IWR) to recognize them. IWR works using the HMM models which were trained by speech databases, for more information about recognizer refer to [8]. In the following table word recognition accuracy is summarized for all 5 systems. Although the results do not seem satisfying, a special feature extraction procedure can improve overall performance of system.

Table6. Word Accuracy of speech databases

Speech Database	Word Accuracy
Original FARSDAT	61.5
Reduced FARSDAT	71.1
Transferred FARSDAT (dialogic to dialogic)	76.1
Transferred FARSDAT (dialogic to modem)	35.9
Tel FARSDAT	52.0s

5. CONCLUSION

In this research, we exhibit an efficient method for converting the direct speech databases to telephony ones. By using this method, the required time for building a new telephony database from an existing direct database is as short as the length of database plus a very short time for preparing speech files.

The main characteristic of this method is the naturalness of resulted voice and its similarity with real telephony signals. Also context independency of this method makes it as a general framework for telephoning any direct speech databases.

Another feature of our framework is the ability of implementation by different hardware devices makes it possible to produce a dedicate database for each special usage which leads to more accurate recognition. Adding this feature to the ordinary telephony databases is impossible.

ACKNOWLEDGEMENT

Special thanks to the Mr. Amir Sanian for his comments and helps for dealing with Dialogic board, Mr. Bagher Babaali for his valuable proposals about method of performing tests and also Mr. Khosro Hossienzadeh and Mr. Mohammad Bahrani for their intuitive comments and useful utilities.

This work could not be done without facilities of Asr Gooyesh© and Jhoobin© companies.

REFERENCES

- [1] M. Bijankhan, et al., "FARSDAT: Farsi Spoken Language Database," *In Proc. of 5th ICSST*, vol. 2, pp. 826-829, Australia.
- [2] J. Sheikhzadegan, M. Bijankhan, "Speech Databases of Persian Language," *In Proc of Second Wookshop on Persian Language and Computer*, Iran, June 2006.
- [3] Intel Dialogic D/4PCIU Voice Board, Datasheet
- [4] Voice Software Reference: Programmer's Guide, 2002 Dialogic Corporation, 05-1456-004.
- [5] Data Transmission System and Equipment – Serial Asynchronous Automatic Dialing and Control, December 1999, PN-4686
- [6] TIA TR-29.2, TIA IS-101, Facsimile Digital Interface – Voice Control Interim Standard for Asynchronous DCE, 1993.
- [7] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *IEEE Trans.* vol. 77, pp. 257–286, February 1989.
- [8] S. Vaisipour, B. Babaali, H. Sameti, "A Rescoring method for detecting Out of Vocabulary words in a phoneme-based IWR," *In Proc. of 13th IWSSIP*, September 2006.