Speech recognition simulation and its application for Wizard-of-Oz experiments

Alex Trutnev, Antoine Ronzenknop, Martin Rajman

Artificial Intelligence Laboratory
Institute of Core Computing Science
School of Computer and Communication Sciences
Swiss Federal Institute of Technology
IN (Ecublens), CH-1015 Lausanne (Switzerland)
{alex.trutnev, antoine.rozenknop, martin.rajman}@epfl.ch

Abstract

This contribution focusses on the simulation of speech recognition engines in the framework of Wizard-of-Oz experiments performed to design and evaluate dialogue-based vocal systems. Such simulation can be useful in cases where different dialogue management strategies are to be evaluated in terms of user satisfaction and where no automatic speech recognition engine or training data are available. The aim of the described work is to build a methodology and a tool that allow to simulate recognition errors in a controlled way. Two approaches are described, that produce, for any set of word sequences, simulated "recognition outputs" (i.e. "noised" versions of the input sequences) in such a way that the obtained average Word Error Rate and Word Accuracy scores correspond as accurately as possible to pre-defined values representative of a targetted "true" speech recognition engine. The first approach aims at simulating Word Accuracy and Word Error Rates levels only. The simplicity of this approach leads to the fact that it does not fully simulate any real speech recognition engine, in the sense that it doesn't produce the same sentences as the real speech recogniser would. The second studied approach integrates the Viterbi decoding algorithm used in many speech recognition systems. The evaluation of the first approach was done using results produced by the Loquendo speech recognition system. The obtained average relative difference is of 1.54%.

1. Introduction

The successfull development of dialogue-based vocal systems directly relies on constant evaluations of its two main components: speech recognition engine (SRE) and especially dialogue management (DM). SRE is relatively independent from the targetted application in the sense that any test data of the same quality can be used for the evaluations; for instance, training and testing on telephone speech extracted from the Swiss French Polyphone database can give reliable measure of recognition performances of SRE used in InfoVox project operating on telephone speech in French (Trutnev and Rajman, 2004; Kommer et al., 2000). The evaluation of the DM, on the contrary, is strongly dependent on the application framework. Even if it is feasible to make offline evaluation of the DM on the basis of registered and transcribed dialogues, it is a hard task to find the dialogues corresponding to the targetted application. Different tests, such as Wizard-of-Oz (WoZ) experiments or field-tests, are thus necessary.

In a WoZ experiment (Fraser and Gilbert, 1991), the designed system is fully or partially replaced by the *wizard*. For example, the wizard can play the role of SRE by manually transcribing the user prompts, or the role of dialogue manager by chosing the actions to perform on the basis of user prompts. One of the main goals of WoZ is the acquisition of data for training-adaptation of acoustic and language models used in speech recognition. The other important aim is evaluation of dialogue management strategies, or, in other words, the evaluation of the impact of using of different dialogue management strategies on the global system performances expressed in terms of user satisfaction (Gauvain et al., 1995; Minker, 1998). From this point of view, we must have a tool that controls speech recognition in order to "play" with its quality, with the Word Accuracy scores varying from 100% to 0%.

As an example of such a tool, the one used in WoZ of the Boris project (Dudda, 2001) can be mentionned: the acquired user prompts are transcribed by the wizard in the specific eletronic form, and then treated by speech simulation resulting in a noised version of the prompts. Four levels of recognition accuracy are implemented: 100%, 83%, 75%, 67%. In the first case, all the informations acquired by wizard are passed to the system. In other cases, in order to achieve desired accuracy, one sixth, one fourth and one third word in acquired prompt are masked so that they cannot be found in the phonetic lexicon (an 'X' character is added at the beginning of the chosen words). This approach is interesting in its implementation simplicity, but nevertheless suffers from the lack of "realism", e.g., speech recognition errors are composed not only of deletions, but also insertions ans substitutions.

This contribution studies more advanced approaches trying to achieve more accurate and realistic simulations. It is organized as follows: section 2. gives more details about the framework of speech recognition simulation; section 3. describes the first proposed approach of speech recognition simulation; section 4. reports about the second approach; the results of the the evalution of the first approach approach are reported in the section 5.; finally, the section 6. provides discussion about the approaches and their evaluations.

2. Speech recognition and its simulation

Speech recognition is a stochastic process characterized by parameters such as acoustic and language models, and *meta-parameters* such as relative importance of acoustic and language models or the word insertion penalty. Each parameter suffers from the *training* problem which is linked to the fact that the trained parameters achieve *local minima*: due to the unavailability of the necessary training

data for most applications, the combination of these "locally" trained parameters results in overall recognition performances worth than those of human interlocutors, especially in the case of unconstrained continuous speech.

The evaluation of SREs is based on Word Accuracy (WA) and Word Error Rate (WER) scores (Gillick, 1989; Pallett, 1990). These scores measure the recognition performances of a given SRE for a given application: WA tells the proportion of well recognized words, WER indicates the proportion of misrecognized words.

The first idea of speech recognition simulation would be thus to directly operate on the WA and WER scores, as done in Boris project, but putting more theoretical sense in the choice of the words to be corrected. This idea is used as basis of the first studied approach, named *Local* approach.

Then, one can question what processes real SREs implement, i.e. what "laws" constrain the SREs to correct the words. The idea here would be to simulate these laws in order to produce words sequences close to the ones that the real SREs would produce. The Viterbi decoding algorithm is used in many SREs. We propose to modify this algorithm, together with the models, to produce the desired simulated word sequence by "decoding" the initial correct one. But in this approach too, and despite the fact that it is much closer to the real SREs than the first approach, one cannot estimate that it produces exactly the same word sequences as real SRE using the same models.

The following two sections describe in details these approaches.

3. Local approach

In this approach, we model the speech recognition as a process performing 4 actions on each word: leave the word unchanged, replace the word, delete the word, insert new word. A probability is associated with each of the actions, estimated on the basis of our expectation of the overall WA and WER. The problem here is therefore to estimate the probabilities.

We first define the following variables:

$WA = \frac{1}{n} OK$

and

$WER = \frac{1}{n} (DEL + INS + SUBST)$

, where OK is the number of well recognized words, DEL is the number of deleted words, INS is the number of inserted words, $SUBST$ is the number of substituted words, and n is the size of the input sequence:

$n = OK + SUBST + DEL$

The size of the output sequence is

$n' = OK + SUBST + INS$

Then, after the necessary transformations, we obtain the estimations of the probabilities:

$$P_{ok} = \frac{\overline{WA}}{\overline{WER} + \overline{WA}} \quad (1)$$

$$P_{ins} = 1 - \frac{1}{\overline{WER} + \overline{WA}} \quad (2)$$

$$P_{del} = 1 - \frac{\overline{\sigma}}{\overline{WER} + \overline{WA}} \quad (3)$$

$$P_{subst} = 1 - P_{ok} - P_{ins} - P_{del} \quad (4)$$

where $\overline{WA}$ and $\overline{WER}$ are average values balanced by the lenghts n_i of the observed word sequences, $\overline{\sigma}$ is the average balanced expansion rate:

$$\overline{WA} = \frac{\sum n_i WA_i}{\sum n_i}, \overline{WER} = \frac{\sum n_i WER_i}{\sum n_i}, \overline{\sigma} = \frac{\sum n_i \sigma_i}{\sum n_i}$$

and σ_i is the expansion rate:

$\sigma_i = \frac{n'_i}{n_i}$

Finally, the algorithm is the following:

1. initialise the values of P_{ok}, P_{ins}, P_{del} and P_{subst} on the basis of 1-4 (we suppose that $\overline{WA}$, $\overline{WER}$ and $\overline{\sigma}$ are initial parameters);
2. put the first word in the sequence as the current word;
3. randomly choose one operation OK, INS, $SUBST$ or DEL according to the probabilities;
4. if the selected operation is:
   - OK : copy the current word to the output sequence; finish if the current word is the last word in the input sequence; put the following word as the current word;
   - $SUBST$: substitute the current word by some word and copy it in the output sequence; the substitution is done either uniformly or on the basis of words confusion matrix (Trutnev and Rajman, to appear); finish if the current word is the last word in the input sequence; put the following word as the current word;
   - DEL : do nothing with the output sequence; finish if the current word is the last word in the input sequence; put the following word as the current word;
   - INS : choose randomly some word and copy it in the output sequence;
5. return to 3

4. Viterbi based approach

The Viterbi algorithm is an aproximation of the Dynamic programming algorithm widely used in speech

recognition systems. It is used at decoding level when trying to find the most probable word sequence for the input acoustic observations, according to the used acoustic and language models:

$$\log P(W|O) = \log P(O|W) + \log P(W)$$

, where W is the sequence of words, O represents the acoustic observations. Probability $P(O|W)$ is estimated by the acoustic model, $P(W)$ is estimated by the language model.

In the case of speech simulation, the input acoustic observations are the correct word sequence that has to be modified, the acoustic model is replaced by the phonemes confusion matrix[1] (Trutnev and Rajman, to appear), the language model is the same as in real speech recognition. The input word sequence is phonetized with the phonetic model, and each phoneme is repeated, typically the number of times it has been seen in training corpus used to train the acoustic model. The columns of the phoneme confusion matrix is then used to replace each repetition by a probability distribution on the phoneme set. After that, the proper standard Viterbi decoding can begin.

We can use meta parameters similar to the ones used in speech recognition, in order to constrain the simulator to produce the desired performance in terms of WA and WER. As the matter of fact, although we can guarantee closeness of the produced word sequences to real recognized ones, it is not predictible what words will be deleted, inserted or substituted. As an example of a meta parameter, a *furge factor* corresponding to the relative weight of acoustic model vs language model can be considered. This parameter can be defined as $\frac{1-\alpha}{2}$ multiplication factor added to $\log P(O|W)$ and $\frac{1+\alpha}{2}$ multiplication factor added to $\log P(W)$. α has to be trained for a given WA and WER. The approach performs as follows: when α is set to -1, i.e. absolute confidence is given to the "acoustic model", then the recognzied sentence is the same as the input one. If α is set to 1, then the output is always the same, and it is the most linguistically probable word sequence.

Of course, we will need another meta parameter in order to fine-tune WA and WER independently from eachother. However, the actual work consists in a first evaluation of the forementionned approach.

5. Evaluation of the local approach

Evaluation of the first (local) approach was undertaken on the basis of data acquired during Inspire project (Rajman et al., 2004). 137 sentences in German pronounced by 10 speakers were first recorded during a preliminary WoZ experiment. Then, 21 "office conditions", corresponding to different noising techniques (Smeele et al., 2004), were applied on that data set. The resulting test data sets were recognized by Loquendo speech recognition system[2] (Trutnev and Rajman, 2004). Evaluations resulted in a set of WA and WER scores. The simulation system was then used to treat 137 sentences with 21 sets of parameters. The obtained noised sentences were compared to the original sentences. The resulting WA and WER scores were compared to Loquendo scores. The relative average difference is of 1.54%.

Table 1 gives an extract of the recognition results.

Conditions	WA	Subst	Del	Ins	WER
01_03a_2	51.8	37.1	11.1	5.0	53.2
noised	51.3	39.3	9.4	4.6	53.3
rel. diff.	0.5	2.1	1.7	0.4	0.1
01_03a_3	41.7	47.3	11.1	18.1	76.4
noised	41.3	52.6	6.1	16.6	75.3
rel. diff.	0.4	5.3	5.0	1.5	1.1
...	...	...	...	...	...
07_09_4	43.6	43.0	13.4	5.6	62.1
noised	43.4	45.6	11.0	4.6	61.2
rel. diff.	0.2	2.6	2.4	1.0	0.9
07_09_c	11.9	69.6	18.4	11.9	99.9
noised	11.2	77.2	11.6	7.1	95.8
rel. diff.	0.7	7.6	6.8	4.7	4.1
...	...	...	...	...	...
17_20_c	36.2	48.9	14.9	6.4	70.1
noised	35.7	51.8	12.4	4.7	69.0
rel. diff.	0.5	2.9	2.5	1.7	1.1
21_24_2	47.5	40.0	12.5	5.2	57.7
noised	47.2	42.7	10.2	4.6	57.4
rel. diff.	0.3	2.7	2.3	0.6	0.3
...	...	...	...	...	...

Table 1: Evaluation of local approach: lines *01_**, *07_**, *17_**, *21_** contain results of recognition of produced with with Loquendo, "noised" stands for the recognition results produced with the local noising approach, "rel. diff." stands for the relative differences between Loquendo and local approach scores.

6. Discussion

The evaluation of the first proposed approach for speech recognition simulation shows its closeness to the real SRE in terms of WA and WER. In addition, it has considerable advantage vs the second approach: its implementation simplicity, since it requires as input data only the desired scores and the original word sequence.

The interest of the second approach is due to the fact that it uses the same techniques as the real SREs. This can be used in the cases when the specific evaluations, e.g. systematic substitution or deletion of given words, is necessary.

One important feature shared by both approaches should be highlighted: the management of substitution cases. It is made on the basis of phonetic proximities between words and phonemes. Correct estimation of the confusion matrices can guarantee that the substitutions com-

[1] Phonemes confusion matrix is a table in which the cell $[i][j]$ corresponds to the probability of confusing the phoneme i with the phoneme j, or, more precisely, to the probability $P(i|j)$ that the phoneme i is recognized when the phoneme j is pronounced.

[2] http://www.loquendo.com/en, commercial speech recognition engine, the used version was 5.9

mitted during the simulation processes are the same as during the real speech recognition.

As an improvement of substitution in the first approach, it should be considered the substitution of one word by a sequence of words, or vice versa. This will have an important impact on the algorithm, and the formulae of estimation of the probabilities should also be modified.

7. References

Dudda, C., 2001. *Evaluierung eines natrlichen Dialogssystems fr Restaurantausknfte*. Ph.D. thesis, Institution fr Kommunikationsakustik, Ruhr-Universitt, Bochum, Deutschland.

Fraser, N. and N. Gilbert, 1991. Simulating speech systems. *Computer Speech and Language*, 3-5.

Gauvain, J-L., S. Bennacef, L. Devillens, S. Lamel, and S. Rosset, 1995. The spoken language component of the mask kiosk. In *Human Comfort and Security Workshop*.

Gillick, Cox. S., L., 1989. Some statistical issues in the comparison of speech recognition algorithms. In *ICASSP*.

Kommer, V. R., M. Rajman, and H. Bourlard, 2000. Heading towards virtual-commerce portals. *Comtec*:10–13.

Minker, W., 1998. Evaluation methodologies for interactive speech systems. In *in Proc. of First International Conference on Language Resources and Evaluation (LREC1998)*.

Pallett, et al., D., 1990. Tools for the analysis of benchmark speech recognition tests. In *ICASSP*, volume 1.

Rajman, M., H. Bui Trung, A. Rajman, F. Seydoux, and A. Trutnev, 2004. Assessing the usability of a dialogue management system designed in the framework of a rapid dialogue prototyping methodology. *Submitted in Acta Acustica*.

Smeele, P., J. Krebber, S. Mller, T. Ganchev, A. Vovos, and B. Kladis, 2004. System component assessment report. Technical Report Deliverable 6.1, IST project INSPIRE (INfotainment management with SPeech Interaction via REmote-microphones and telephone interfaces, IST-2001-32746).

Trutnev, A. and M. Rajman, 2004. Comparative evaluations in the domain of automatic speech recognition. In *International Conference on Language Resources and Evaluation (LREC2004)*.

Trutnev, A. and M. Rajman, to appear. Enhanced phonetic model for speech recognition.