EFFICIENT ASSESSMENT OF ASR SYSTEMS
BY USING SUBSETS OF A TEST DATABASE

Arkadiusz Nagórski [1,2], Lou Boves [2], Herman Steeneken [1]

[1] TNO Human Factors, Soesterberg, the Netherlands
[2] Department of Language and Speech, University of Nijmegen, the Netherlands
email: nagorski@tm.tno.nl, boves@let.kun.nl, herman@steeneken.com

ABSTRACT

In this paper, assessment of ASR systems with a limited set of speech data selected from a larger testing corpus was studied for connected Dutch digits. Three methods of data selection were applied, namely random, knowledge-based, and data-driven selection. The goal of this study was to find out whether reliable assessment of speech recognition systems can be achieved by using a small sample of the testing corpus. The results are presented in terms of the confidence interval of the mean value calculated for the recognition scores. It appeared that the method of data selection used in this experiment did not contribute significantly to minimize the range of the confidence interval with respect to random selection. Thus, for the speech material presented here, random selection can be successfully applied to obtain a satisfactory assessment even with relatively small subsets of the testing corpus.

1. INTRODUCTION

Assessment of ASR systems is performed with testing corpora that are representative for the application. The testing corpus is usually predefined as an independent part of speech material in a database after the speech data were collected and transcribed. Most of the speech databases that have been collected during the last decade come with a default division in a training, development and testing subcorpus [1]. Corpora for the development of general purpose ASR systems can be further subdivided according to specific recognition tasks, such as connected digit recognition, command word recognition, etc. Even if the ASR system or evaluation procedure is task specific, some general guidelines should apply when creating the testing speech corpora [2]. It is very important that the data in the testing corpus are a representative sample of the population. This is especially important if a general purpose database is used for the development of an ASR system for a highly specific population (e.g. only middle-aged men), or for a very specific task (e.g., command word recognition in helicopters), that may not be very well represented in the training, development and testing parts of that corpus.

The simplest way of being maximally representative, is to obtain a large amount of speech recordings from many speakers in a database that covers a wide range of variability in speech production and recording conditions. The variability in speech can be a result of fixed as well as random factors. Some factors that cause variation are know a priori, and therefore the design of the database can be made to incorporate these factors. Nevertheless, it is difficult to avoid uncontrolled variables from affecting the process of speech collection such as variation in pronunciation and speaking rate, channel distortions, noises or other unknown factors.

When a database is collected for testing purposes, the random variability in the speech data can affect the result, but this usually is not a big concern if the database is large enough. Testing data are representative if they approximate the speech conditions that will arise in real system operating. But if we are interested in a reliable assessment of the ASR system using only a part of the testing corpus, or when only a small testing corpus can be collected, then we should care not only about the data properties that are already known but also pay attention to the variability caused by unplanned factors (whether random or systematic). Last but not least, insight in the impact of planned and unplanned factors on the outcome of the performance evaluation can have substantial diagnostic value.

In this paper we study whether reliable evaluation of an ASR system is possible using a subset of data selected from a larger testing corpus. In this way we can reduce the amount of speech data that are necessary to check the recognition performance and at the same time get the assessment result very close to what we would have obtained for all the data in the testing corpus. To investigate this issue, we compared three methods of data selection: random (rand), knowledge-based (kb), and data-driven (dd) selection. The first method serves as a point of reference for the other two methods. Random selection is the simplest manner of data selection one can apply. The knowledge-based or data-driven selection uses some knowledge about speech data in the selection of the testing set. This knowledge is derived from properties of speech material described in the database (kb) or from the analysis of speech data in terms of their acoustic features (dd). The known speech properties determine the planned variability in speech, while the acoustic analysis can detect the variability in speech data caused by unplanned factors.

2. SPEECH DATA AND ASR SYSTEM

In this work we continued experiments with speech data from the Dutch Polyphone database limited to connected digits [3]. The same training and testing corpora were used in the earlier studies [4, 5]. The corpora were chosen arbitrarily from all the data available in the database and contained 78344 phones (training) and 64320 phones (testing) in total. The average length of utterances (recordings) was 20.89 phones for the training corpus and 20.76 phones in case of the testing corpus.

The ASR system was built using the HTK Toolkit version 3.0 [6]. To cover the phones in Dutch digits, we had 18

phone-level HMMs with 3 states and left-to-right topology. Additionally, two models were created to deal with silence /sil/ and noise /noi/ [5]. The ASR system was trained on all the training data and configured to perform unconstrained phone recognition because we are mainly interested in the impact of the quality of the acoustic models on the recognition performance and wanted to minimize the impact of the language model. Scoring was performed in accordance with the NIST algorithm (weights 3, 3, 4) to match the recognized phonetic transcription with the reference one [6].

Speech data were parameterized into 16 MFCCs (c0-15) from 16 band Mel-filter spectra in the frequency range from 80 Hz to 3800 Hz, calculated every 10 ms from 16 ms Hamming windowed frames. Cepstral mean normalization and liftering were also applied during the feature extraction.

3. SELECTION ALGORITHM

For the work presented here we developed a selection algorithm that finds a subset of utterances (recordings) from the corpus that fulfils a number of specific criteria. The criteria relate to the size of the set and to the distribution of specific properties of the speech data. The relevant properties are described below for the two active methods of selection.

The selection algorithm works in two phases. The first phase consists of a random selection, where only the size of subset is controlled in terms of the total number of phones. The second phase is intended to optimize the contents of the random set so that the criteria requested are satisfied as much as possible. The optimization procedure used in the second phase is identical for the knowledge-based and data-driven selection, but the optimization criteria differ between these two treatments. In this phase both the size of data set and the distribution of properties are controlled. In this study, the distributions optimized during the second phase of the selection algorithm are referred to the corresponding statistics collected for all the data in the testing corpus (population) since, as mentioned earlier, we are interested in the assessment of the ASR system with a subset of data that would give a reliable estimate of the results obtained for the whole testing corpus. However, the procedures that we have implemented can use any reference in the selection process, depending on the experimental aims. The first phase of selection algorithm was introduced in order to randomize the starting point of the second phase that essentially implements a hill climbing procedure. In this way we can obtain many subsets of similar size having different content. Also, we can obtain estimates of the statistical stability of the properties of the sets.

The active phase of selection methods is based on the greedy search approach [7]. The algorithm implemented in this experiment performs alternate backward-forward greedy search. In each step of the search, the current distribution in a selection is compared with the reference distribution in order to determine which utterance should be added to or removed from the current selection to minimize the discrepancy between the distributions. The inclusion/removal criterion is based on the Root Mean Square Error (RMSE) between the current distribution **d** and the reference distribution **d**$_{ref}$ calculated for each property p, where **d** denotes a vector and N the number of elements in the vector:

$$RMSE_p = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{d}(i) - \mathbf{d}_{ref}(i)\right)^2} \qquad (1)$$

Next, the RMSE values obtained for each property are weighted to calculate the total distribution error. The weighting is introduced to make the search algorithm sensitive to the scale of the distribution error contributed by the individual properties during the selection process. The weights are adapted dynamically in every step of the search and are proportional to the distribution error in the corresponding properties observed in the previous step. Due to this weighting scheme, the search algorithm becomes less sensitive to the properties for which the current distribution already matches the reference one very closely (small error) and optimizes those properties for which the discrepancy between the distributions is relatively larger (big error). The search process stops when the same utterances are repeatedly added and removed, in which case the distribution error cannot be further minimized.

4. KNOWLEDGE-BASED SELECTION

The knowledge-based selection presented in this work made use of the known (identified) properties of speech data that can be derived from the meta-data in the database. Based on an analysis of the information provided with the Polyphone database, we decided to focus on the following speech data properties in the knowledge-based selection:

- property 1 (2 values): gender of speaker (female or male),
- property 2 (2 values): age of speaker (range 21-40 or 41-60 years old),
- property 3 (12 values): dialect or regional background of speaker (12 provinces),
- property 4 (50 values): utterance length (from 2 up to 58 phones per utterance),
- property 5 (18 values): occurrence of phones (18 phones).

For each of these five properties, the appropriate histograms where collected from all the data present in the testing corpus and the vectors with reference distribution were created.

5. DATA-DRIVEN SELECTION

The data-driven selection operates on acoustic parameters of the speech data in order to get the requested result of selection. Similar to the knowledge-based approach, we are interested in finding subsets of utterances that will approximate a specific distribution of speech data properties. But different from the knowledge-based approach, here the properties are primarily related to acoustic features and not to the categories that are already identified such as gender, age or dialect of speakers. The data-driven selection investigated in this experiment is based on two properties:

- property 1 (360 values): occurrence of phone observations in classes of acoustic features (360 classes)[1],
- property 2 (50 values): utterance length (from 2 up to 58 phones per utterance).

---

[1] The total number of classes is also the result of a data driven procedure that will be explained below.

The first property is fully specified for the data-driven approach, but the use of the second property (utterance length) seems to be debatable since it can be presented both in the knowledge-based and data-driven approach. Nevertheless, utterance length appears to be an important factor to control in selection of speech data as shown in the earlier work [5]. Therefore, we decided not to ignore this property, as the recognition performance needs to be checked due to a particular choice of testing data.

In order to find the classes in the space of acoustic parameters, speech data from the testing corpus were processed and analyzed similarly to the method presented in [5]. Both the ASR system and data-driven selection worked with the same kind of features (MFCC). Data observations (supervectors) used to characterize speech sounds were created for every phone present in the testing corpus. A single-phone supervector holds data of 3 states of a token of a phone with 16 MFCCs (c0-15) averaged over the duration of each state (the state occupancy of individual frames was determined by the HTK recognizer). The total number of variables in a supervector was thus equal to 3 x 16 = 48. Next, Principal Component Analysis (PCA, [8]) and clustering (K-Means) were performed in each of 18 phone spaces separately. In the current experiment, the PCA and clustering settings were different from those used in [5]. The original 48 dimensions of the supervectors were reduced to 16 Principal Components (PCs). Thus, we ended up with vectors with the same number of elements as in a single frame.

The clustering was performed in 16 PCs of each phone space. The process was terminated when the smallest resulting cluster would contain fewer than 65 observations/tokens. Thus, the reduction factor of selection was set to RF=65 what theoretically gives a possibility to reduce the size of the testing corpus to approximately 1000 phones (64320/65) while retaining at least one phone observation in the smallest cluster(s) found. Nevertheless, empty clusters cannot be avoided especially when a large number of conditions must be simultaneously satisfied during the search. Moreover, the speech material in the database was organized in the form of utterances, and only complete utterances could be included or removed in the optimization of the testing sets. Therefore, the selection process could not optimize the sets by adding or removing individual phone tokens.

Similarly to the case of knowledge-based selection, the reference distributions used in the process of data-driven selection were obtained from all data in the testing corpus.

6. EXPERIMENT

Using each of the selection methods investigated in this experiment, we selected data sets of 75 fixed sizes in 30 trials from the testing corpus. The requested size of the testing set was chosen in logarithmic steps and ranged from 100 up to 59566 phones. Due to the fact that utterances contained different number of phones, the size of selection could vary from the requested one, especially for the smallest selections.

We used several measures to compare the performance of ASR system and the selection methods depending on the size of testing subsets. In this paper we present two of them, namely the confidence interval for the mean value of the phone deletions (D), substitutions (S) and insertions (I), and the recurrence level of selection.

The confidence interval was calculated for a significance level of 5 %. For the calculation of the confidence interval the Gaussian distribution was used, which is justified given the relatively large number of trials that were performed (30 trials).

The parameter called 'recurrence level' was introduced as a measure of the selection recurrence. In this experiment, we assessed the recognition performance of the ASR system in a number of random trials to get average results. Thus we expect that the content of the testing sets (utterances) were random as well. However, the utterances were selected from the testing corpus of limited size and according to the specific conditions. Especially in the case of knowledge-based or data-driven selection it could happen that the selection algorithm was repeatedly compelled to add a part of the same utterances to the set in order to fulfil the search conditions. To obtain insight in the proportion of recurrent data selected in the sets, we compared their contents. We made such a check between sets obtained for the same selection method (self-test) and for different selection methods (between-test).

7. RESULTS AND DISCUSSION

Figure 1 presents the confidence interval for the mean value of D, S and I score calculated for phones depending on the selection method and size of the testing set selected. Before we start to analyze the results presented in Figure 1, it is worth to discuss the recurrence level of selection and performance of the search algorithm.

It appeared that the recurrence level of selection was strongly dependent on the size of selection. For sets with up to approximately 6000 phones, there was on average less than 10 % phones recurrent in the selection trials (both for the self and between-test). Nevertheless, it happened that some of the smallest sets had the recurrence level at 45-55 % (maximum observed). But such a coincidence was very unlikely, since the mean value of recurrence level retained its strong downward tendency when the size of the testing set was decreasing. Thus, in case of the testing corpus we investigated, the upper limit of 6000 phones seems to be a reasonable threshold to treat the selection trials as yielding independent data sets. For bigger testing sets, the average recurrence level increased relatively fast. Therefore, it is difficult to consider these selections as mutually independent. Moreover, we also noticed that each set bigger than 2900 phones had a recurrent part of the content. The similarity of tendencies for the recurrence level observed between different selection methods (rand, kb, dd) as well as the type of comparison (self or between) suggests that these dependencies are specific for the size rather than for the methods of selection that were investigated.

The search algorithm for knowledge-based and data driven selection noticeably reduced the initial discrepancy in distribution of the properties typical for the random selection (first phase). Due to the search process, the distribution error decreased on average 7.4 times in case of the knowledge-based and 3.5 times in case of the data-driven selection. The optimization was more effective when the size of selection was increasing, up to the moment when nearly half of the data from the corpus were selected.

We focused on analysis of the confidence interval since we expected to observe a change in its range for different selection methods. A smaller confidence interval observed for one of these methods would imply that the range in which we
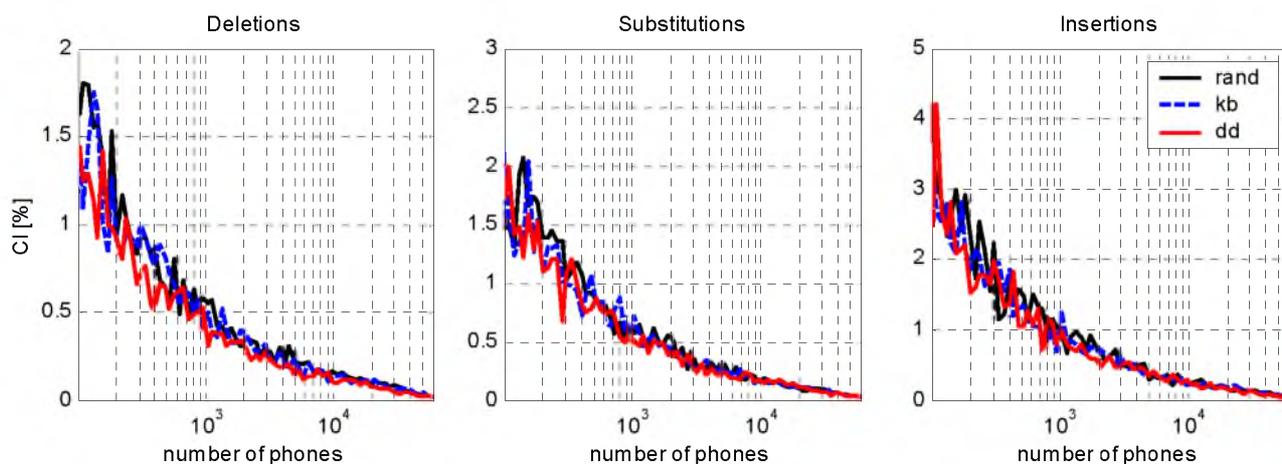
Figure 1. The confidence interval (*CI*) for the mean value of the phone deletions (*D*), substitutions (*S*), and insertions (*I*) in percentages [%] obtained in the case of random (rand), knowledge-based (kb) and data-driven (dd) selection as a function of the size of the testing set. Note that the confidence interval presented in this figure is 1-sided and centered to the estimate of mean values.

can expect the true result became narrower. According to what we found about the recurrence level of selection, we should focus on analysis of the results presented in Figure 1 for the data sets smaller than 6000 phones. In this range, we can observe that the confidence interval for the mean value of *D*, *S* and *I* score is rather similar for all the three selection methods. However, we noticed a slight advantage in case of the data-driven selection (smaller confidence interval). This advantage is most evident for the testing sets smaller than approx. 3000 phones, especially in terms of deletions. Nevertheless, the results do not suggest that active selection of testing data will allow one to reduce the size of the testing set without compromising the reliability of the results.

The failure of active selection to reach the asymptotic test reliability faster than random selection is probably due to the similarity between the statistical distributions of the testing and training corpus. A random selection of sufficient size from the testing corpus will then tend to approximate the distribution of the corpora so closely that the added value of clever selection techniques is negligible. Yet, we expect that we will gain substantial diagnostic information by turning the selection process upside down, and search for utterances that are likely to be outliers, rather than middle of the road.

The fact that we applied active selection only to a database of connected digits, where the variability in the properties of the speech is inherently limited, has probably contributed to the failure of active selection to outperform random selection. In future work we will apply our approach to selection of testing data to speech databases with a larger vocabulary and different speech styles (read or conversational).

8. CONCLUSION

The comparison of three selection methods showed that the random, knowledge-based and data-driven selection of speech data from the testing corpus result in a similar confidence level for the assessment of recognition performance of the ASR system. Although we can observe some advantages of the

knowledge-based and especially data-driven over the random selection, the results are not significantly different.

The analysis of recognition performance in combination with the recurrence level suggests that a reliable evaluation of ASR systems can be achieved starting with subsets bigger than 1/10 part of the present testing corpus. In this situation, the random selection of data is satisfactory, since the probability that a subset is strongly biased towards utterances that are extremely easy or difficult to recognize decreases as the size of the sets increases.

REFERENCES

[1] The SpeechDat Projects, http://www.speechdat.org

[2] EAGLES, "Handbook of Standards and Resources for Spoken Language Systems", *Walter de Gruyter Publishers*, Berlin & New York, 1997.

[3] E.A. den Os, T.I. Boogaart, L. Boves, and E. Klabbers, "The Dutch Polyphone Corpus", *Proceedings Eurospeech*, Madrid, vol. 1, pp. 825-828, 1995.

[4] A. Nagórski, L. Boves, and H. Steeneken, "Optimal Selection of Speech Data for Automatic Speech Recognition Systems", *Proceedings ICSLP*, Denver, Colorado, vol. 4, pp. 2473-2476, 2002.

[5] A. Nagórski, L. Boves, and H. Steeneken, "In Search of Optimal Data Selection for Training of Automatic Speech Recognition Systems", *Proceedings IEEE ASRU*, St. Thomas, U.S.V.I., pp. 67-72, 2003.

[6] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK Book ver. 3.0", *Cambridge University*, 2000.

[7] T.H. Cormen, C.E. Leiserson, and R.L. Rivest, "Introduction to Algorithms", *The MIT Press*, London, 1990.

[8] I.T. Jolliffe, "Principal Component Analysis", *Springer-Verlag*, Berlin, 1986.