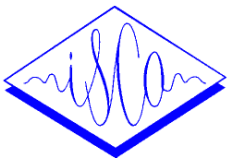

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Telephone Speech Recognition from Large Lists of Czech Words

Jan Nouza

Technical University of Liberec, Department of Electronics and Signal Processing,
Halkova 5, 461 17 Liberec, Czechia
jan.nouza@vslib.cz

ABSTRACT

In the paper we investigate methods suitable for practical implementation in a recognition system that is to classify telephone input in form of isolated words/phrases belonging to large vocabularies with equiprobable entries, such as people names, city and local names, etc. Specifically for Czech language we propose a pronunciation lexicon with a prefix-stem-sufix arrangement combined with appropriate caching and pruning techniques and a 2-level (monophone and triphone) based classification. In experiments done with telephone speech containing items from a 5347-word city-name vocabulary we obtained 90.1 % recognition score in average time 645 ms per word. Acoustic models for these experiments have been trained on an only available multi-speaker database that was originally recorded by a microphone and later transferred over telephone lines and automatically realigned.

1. INTRODUCTION

In many recent automated telephone systems providing information, reservation or a various kind of transaction services, the classic hear-menu-touch-button scheme has been replaced by a more natural both-side voice communication. The complexity of such systems may vary from simple ones operating with a small list of keywords to very sophisticated demonstrators enabling almost a natural dialogue between a user and the remote computer [1]. The experience with practical performance of these services (e.g. [2]) indicate, however, that at the current level of speech and language technology the most elaborated systems do not necessarily achieve the expected high success rate when opened for a wide public. In many application fields, a clearly structured system-controlled dialogue scenario seems to be more appropriate, especially for novice and occasional users.

We have arrived at the same conclusion with our system called InfoCity [3]. It is the first fully voice-operated telephone application developed and run in Czechia. It has been in public use since 1999 and it has done a pioneering work when introducing this type of service to Czech people. The response from the users showed that they would have appreciated a similar access to other sources of information, like telephone and zip code directories, complete train and coach time-tables or city transport navigation. All these tasks may be implemented within the existing structured-scenario platform, but they assume vocabularies with hundreds to several thousands of words that must be, in most cases, searched through in an exhaustive way, without relying on any syntax or grammar.

In order to develop and implement an isolated-word/phrase recognition system for large vocabularies of Czech words, we have tried and evaluated several different approaches. Most of them are described in this paper.

2. RECOGNITION GOALS AND SYSTEM SPECIFICATIONS

Our primary goal was to develop suitable methods for real-time recognition of short Czech utterances (single words or phrases considered as single entries) that are spoken in isolated way as responses to systems prompts. We assume vocabularies like, lists of names, surnames, local names, company names, parts of dictionaries, etc. The size of these lists may exceed 10K items, each of them being equally probable. Further, we need to deal speech that is automatically detected in signal coming from analog telephone line. The combination of telephone signal quality and automatic speech detection makes the task even more complicated, because due to various factors (line noise, background noise, barge-in) the utterance endpoints may not be set correctly and the classifier must cope with over-sized (which is better) or partly incomplete (the worse case) speech input.

The main priority of the design is the recognition accuracy achieved under field and, in particular, real-time conditions. In case of smooth telephone transactions the maximum tolerated delay caused by the classifier should not be longer than one second. The complete recognizer is supposed to run on a standard PC equipped by a telephone board. Because there is a frequent demand to handle several (1 - 8) telephone lines simultaneously, we also should consider memory requirements of the proposed methods.

2.1. Evaluation conditions

For evaluation purposes we have collected four test databases based on the following vocabularies:

1. Names - 360 Czech first names,
2. Surnames - 800 selected Czech surnames,
3. City names - 5346 local names,
4. Voc10k - 10 000 most frequent Czech words.

More details about the vocabularies and their complexity can be found in [4]. Each test database consists of 2000 items recorded by 10 speakers over 2 different phone (fixed) lines and captured in the same way as in the target telephone recognition system.

3. SPEECH PROCESSING AND MODELING

Here we briefly describe the speech processing procedures that are common for all the methods investigated in the following sections.

3.1. Signal parameterization

The signal received from the telephone line is sampled at 8 kHz/ 16 bit rate and processed in frames 20 ms long every 10 ms. The frame vector consists of 20 features, from which the first 18 are employed for the classification. It is 8 LP cepstral and 8 delta coefficients, delta and delta-delta log energy. The static parameters, biased due the microphone and line characteristics, are normalized over the whole captured utterance by means of the well-known CMS (cepstral mean subtraction) technique.

3.2. Endpoint detection

The presence of the speech in the analyzed signal is detected using the log energy and measuring the relative change of spectrum (the last two features). Trigger thresholds are continuously updated from values belonging to the varying background noise. In order to eliminate at least some errors caused by the detector, the frame sequence supposed to be speech is augmented by including also a certain number of frames (typically 10) before and after the automatically determined endpoints. These extra frames contain either noise and then they are covered by a noise model or they add a part of speech that would be otherwise missing.

3.3. Phoneme and word models

The basic recognition units are phonemes. For Czech language we use the set containing 41 phonemes as they were defined in [5]. In our system each phoneme (including silence and background noise) is modeled by a three state left-to-right HMM. All vocabulary items, more precisely their phonetic transcriptions, are represented by a linear concatenation of corresponding phoneme models.

The models of Czech phonemes have been trained on a database that contains about 15 hours of speech (specially constructed phonetically rich sentences) recorded through various types of microphones by 70 speakers [6]. The database, that had been annotated and segmented on a phoneme level, served for training both context-independent (monophone) and context-dependent (triphone) models. However, when tested on the evaluation databases (those mentioned in section 2.1), the triphone models yielded only a minor improvement (not more than 1 %) paid by an enormous increase in computation time. This will be further discussed in section 4.1.

3.4 Telephone speech models

Unfortunately, the above models are not appropriate for telephone speech recognition, even if the CMS technique is employed. To get better models, first we had to fit the training data to the target environment. We did it simply by transferring the whole annotated database through a phone line. The data was replayed, sentence by sentence, from a loudspeaker to an adjacent telephone microphone and sent through lines and switchboards to the computer with a telephone board. In this way we obtained recordings that had most characteristics of true telephone speech. We repeated this transfer for two types of phone sets and two different paths to simulate a larger variability in channel transfer conditions. Automatic phonetic segmentation of the new signal has been accomplished using correlation with the previously labeled original signal. After training a new set of models, the recognition results improved significantly and yielded a level that was only 1 - 3 % lower compared to the tests done with the same vocabulary but microphone speech.

4. DESIGNING A CLASSIFIER FOR A LARGE VOCABULARY

In this section we consider various arrangements and different parameter settings of a classifier that should be appropriate for the recognition from a large list of Czech words. When comparing different approaches we will aim at getting best recognition rates in minimum time. Within this section we will use mostly the City name database as the benchmarking data. It contains 5346 local names, which is the complete list of Czech cities and villages. The length of the items in this list ranges from 2 to 34, with the mean being 8.7 phonemes per word. All the experiments were conducted on a PC running at 600 MHz.

4.1. Choosing an appropriate model type

The main parameters that influence the recognition rate are the type and size of the phoneme models. We have trained both the monophone and triphone models, using 8, 16 and 32 gaussian mixtures per state. Employing these models in the tests done with the City name evaluation database we obtained values that are summarized in Table 1. We can notice a) the number of mixtures plays a more significant role in case of monophones and b) for the higher number of mixtures the triphones outperform the monophones in an almost negligible way.

Table 1. - Recognition results obtained for different phoneme model types in the City name (5346 word) task.

Model type	Recog. rate [%]
monophones, 8 mixtures	86.7
monophones, 16 mixtures	88.9
monophones, 32 mixtures	89.8
triphones, 8 mixtures	88.8
triphones, 16 mixtures	89.6
triphones, 32 mixtures	90.1

4.2. Likelihood values caching

While the previous subsection dealt with the score, now we will focus on the computation time. From this point of view the most critical issue is the calculation of state output likelihoods, especially in case of larger numbers of mixtures. Fortunately, in a phoneme based system the need for these calculations can be drastically reduced by caching the already computed values and reusing them if the same frame is matched to the same state. The frequency of repeated calculations for the same frame-state pair is very high for monophones, because there are only 41 x 3 different states. Moreover, the cache hit rate increases with increasing the vocabulary. For illustration, in case of the City name vocabulary, 99.8 % calls for the likelihood evaluations can be served by the cache. Obviously, this rate is much lower for triphones, where we have 3810 states. The impact of the caching scheme on the computation time is demonstrated in Table 2.

Table 2. - Comparison of recognition times for the best models from Table 1 in case the likelihood cache is employed. (At this stage no other optimization techniques are considered.)

Model type	Rec. rate [%]	Aver. recog. time
monophones, 32 mix.	89.8	3 020 ms
triphones, 32 mix.	90.1	28 910 ms

4.3. Lexicon organization

4.3.1 Linear lexicon

All the above described experiments were conducted in a very conservative way, using a linear lexicon (each word represented by a complete sequence of its phonemes) and a serial method of classification (word after word). This approach is easy for direct implementation and thus it may serve for the purpose of checking the performance and score values obtained by more sophisticated methods.

4.3.2 Tree lexicon

The idea of arranging a pronunciation lexicon into a tree form is old [7], however, it has been used in most recent speech recognition systems. It not only reduces the memory needed for storing the vocabulary but it perfectly fits with the search methods based on dynamic programming (DP). It actually introduces another level of caching, because the same prefix part belonging to different words can be processed only once.

The concept of the shared initial parts is applicable both to the serial and parallel classification scheme. In the former case it may lead to the strategy described in [4]. The vocabulary is arranged according to the alphabetic order so that the adjacent words share larger or smaller areas in the frame-state plane, which reduces the space to be searched by the DP algorithm. In the latter approach the words are processed in parallel, frame by frame. Generally, this requires a larger memory space and some overhead operations when compared with the serial classifier. However, if the Viterbi search computation is well organized, we may suffice with only one vector of accumulated likelihoods, in which old values (those from the previous frame evaluation) are continuously updated by new values in a bottom-top direction. However, the most important advantage of the parallel processing is the possibility to introduce pruning techniques.

4.3.3 Prefix-stem-suffix lexicon

The Czech language is known for its complex morphology, based on the frequent use of prefixes and suffixes. This fact complicates the recognition in a significant way, because the common language inventory exceeds half a million different word forms. On the other side, the existence of a limited set of prefixes and suffixes can be used for a more efficient coding of the recognition vocabulary.

Since the prefixes are already captured in the tree arrangement, we should try to find a way how to handle the suffixes. (Here, by the term suffix we mean the frequently occurring final parts of Czech words, no matter whether they carry morphological meaning or not.) To illustrate the existence of the frequent suffixes, let us use

examples taken from the vocabularies of the four evaluation databases:

- In the 360-word Names list, almost one third of the items end by syllables *-na, -la, -ta, -ka* or *-da*, 32 names include trailing parts *-slav* or *-slava*, etc.
- The Surnames set exhibits a feature, that is unique for Czech: female surnames. They differ from the male ones just by suffixes, the most frequent one being *-ova.*
- In the City name list, almost 35 % names end by *-ice* or *-ovice*, 12% ends by *-ov*, and other 8 % end by *-in*, or *-any*.
- The Voc10k list is quite specific, because among the 10K items, there are just about 4K different lexical words while the rest are morphological derivatives of the former.

In our experiments the suffixes have been derived on the statistic rather than morphologic base. For each vocabulary we found the final word parts that were at least two-phoneme long and occurred more than five times. The separation of the suffixes led to another reduction of the search space, similar to the prefix based tree arrangement. This is illustrated in Table 3.

Unlike the prefixes, the suffixes need a special care during the classification. They are handled in the way similar to connected word recognition. In practice it means that for each frame the initial parts of all the words are evaluated first and then the best hypotheses are passed to the corresponding suffix states. If it is done in the most straightforward way it leads to the one-best winner for each suffix. If *N*-best candidates are needed, e.g. for a potential second match, more instances of the most likely suffixes can be generated dynamically. The impact of using the suffixes on the classification time reduction is demonstrated in Table 4. In fact, it is not as significant as for prefixes, because the procedure requires some overhead computation. Moreover, its significance further decreases when pruning techniques are applied. Unlike them, however, the use of prefixes and suffixes does not introduce any loss of recognition accuracy.

Table 3. - Comparison of the total number of states that must be considered (in an exhaustive search) for three different lexicon arrangements.

Total number of states for different types of lexicon			
Vocabulary	**Lin. lexicon**	**Tree lexicon**	**Prefix & suffix**
Names	9 322	5 791	3 563
Surnames	22 833	15 099	8 737
City names	176 785	102 133	53 743
Voc10K	278 546	115 487	47 989

4.4. Pruning

Pruning, i.e. cutting out the less likely hypothesis, is the most efficient way of the computation load reduction. On the other side, it is an heuristic approach that may have a negative impact on the recognition. For example, in case the word start was wrongly determined and shifted to the right, early hypothesis, that would eventually lead to the correct result, could be omitted. For this reason we were very conservative in setting the pruning threshold. The rule we applied was the pruning must not - in any single case - influence the score of the top *N* candidates. (We have chosen *N* to be 5.) This rule safes the accuracy and allows us to do an optional second-level match with these *N* candidates. Moreover, besides the primary state-based pruning we introduce also a similar treatment on the word level. A word hypothesis is taken out of consideration if all word states have been pruned off. The time reduction effect of the pruning schemes is summarized in Table 4.

Table 4. - Comparing various computation reduction schemes in the City name task

Computation reduction technique used	**Aver. recog. time [ms]**
Serial classification, tree lexicon	1 250
Parallel classification, tree lexicon	2 063
Parallel, prefix & suffix	1 642
Parallel, prefix & suffix, state pruning	1 165
Parallel, prefix & suffix, state & word prun.	620

4.5. Two-level classification

If the classification procedure is arranged so that it generates an *N*-best list, we can utilize it in a second level match. Since the *N* is typically in range 3 to 10, the computation load of the method used at that second level does not play a significant role. In our system this final match is based on the triphone models.

5. EXPERIMENT RESULTS

In this section we just summarize the results that were achieved with the four evaluation databases. The results correspond to the optimal setting of the recognition system, that was same for all four databases. It included an automatically derived prefix & suffix lexicon, state & word pruning, 32-mixture monophones at the first level

and 32-mixture triphones at the second level 5-candidate match. The results are summarized in Table 5.

Table 5. - Recognition rates and times achieved with the best classifier arrangement for the four evaluation databases

Database	Rec. rate [%]	Aver. recog. time [ms]
Names	92.3	210
Surnames	91.2	386
City names	90.1	645
Voc10K	79.6	748

5. CONCLUSIONS

In this paper we investigated methods that are suitable for real-time isolated word recognition form large lists of Czech words, in the special case when no syntax nor grammar is applicable. We show that with the combination of several different caching schemes, a specific lexicon arrangement that takes into account frequent prefixes and suffixes, and a two-level approach we can get good recognition results within time less than 1 s even for vocabularies that have more than 10 000 items. We should stress once again that the results summarized in Table 5 were achieved with automatically detected telephone speech. The recognition rate was significantly worse only in case of the Voc10k list. It is because this vocabulary is extremely difficult for recognition since it contains groups of items that are phonetically very similar being morphologic derivatives of the same word (like e.g. *volat, volal, volala, volalo, volali*, etc.).

The proposed methods are not dependent on the acoustic models and that is why we believe that the recognition accuracy may increase if we try to employ more-dimensional feature vectors (e.g. the frequently used 39 MFCCs) and manage to train a better set of triphone models.

Acknowledgments.

The research reported in this paper was supported by the Grant Agency of the Czech Republic (grant no.102/96/KO87), by the Czech Ministry of Education, through research goal project MSM 242200001 and by project Kontakt no. ME293.

REFERENCES

1. Proceedings of VOTS'2000 (Voice Operated Telecom Services) Workshop, Gent, May 2000.
2. Den Os E., Bouves L.: Usability of Automatic Speech Recognition in Telecommunication Services. Proc. VOTS (Voice Operated Telecom Services) Workshop, Gent, May 2000, pp.47-50.
3. Nouza J., Holada M.: A Voice-Operated Multi-Domain Telephone Information System. Proc. of ICASSP 2000, Istanbul, June 2000, vol.VI, pp.3755-3758
4. Nouza J.: A Czech Large Vocabulary Recognition System for Real-Time Applications. Proc. of TSD (Text, Speech and Dialogue) Workshop, Brno, Sept. 2000.
5. Nouza J., Psutka J., Uhlir J.: Phonetic Alphabet for Speech Recognition of Czech. Radioengineering, vol.6, no.4, Dec.1997, pp.16-20.
6. Nouza J., Myslivec M.: Creating and Annotating Speech Database for Continuous Speech Recognition. Proc. of. 4th ECMS Workshop, Liberec, May 1999, pp.147-151.
7. Klovstad J.M., Mondsheim L.F.: The Caspers linguistic analysis system. IEEE Trans. on ASSP, vol. 23, Feb. 1975, pp. 41-46.