INFORMATICA, 2004, Vol. 15, No. 3, 303–314

Limited-Vocabulary Estonian Continuous Speech Recognition System using Hidden Markov Models

Tanel ALUMÄE, Leo VÕHANDU
Tallinn Technical University
Ehitajate tee 5, 19086 Tallinn, Estonia
e-mail: tanel.alumae@phon.ioc.ee, leov@staff.ttu.ee

Received: February 2004

Abstract. The article presents a limited-vocabulary speaker independent continuous Estonian speech recognition system based on hidden Markov models. The system is trained using an annotated Estonian speech database of 60 speakers, approximately 4 hours in duration. Words are modelled using clustered triphones with multiple Gaussian mixture components. The system is evaluated using a number recognition task and a simple medium-vocabulary recognition task. The system performance is explored by employing acoustic models of increasing complexity. The number recognizer achieves an accuracy of 97%. The medium-vocabulary system recognizes 82.9% words correctly if operating in real time. The correctness increases to 90.6% if real-time requirement is discarded.
Key words: continuous speech recognition, hidden Markov models, Estonian.

1. Introduction

Speech recognition is the process of turning human speech to the corresponding sequence of words. In the last decades, there have been great advances in that area. For languages like English, many successful speech recognition systems have been developed and commercial systems are widely available.

In spite of active research in the area of phonetics and computational linguistics, there have been few attempts in the area of speech recognition for the Estonian language. However, the first research results about acoustic analysis of Estonian vowel and consonant system and prosody date back to 1960s (Lehiste, 1966). In 1980s, experiments on Estonian vowel recognition were made. In the middle of 1990s, neural nets were used for diphone recognition experiments (Meister, 2001b). In 2000, a prototype for isolated word recognition (Estonian numbers and letters) was developed in co-operation with Institute of Engineering Cybernetics of Minsk (Meister, 2001).

For comparison, the state of speech recognition of other Baltic languages is also in the very first stage. While there are no reports available for Latvian, the Lithuanian related attempts are similar with those on Estonian: there are works on phonetic characteristics of Lithuanian, as well as a few attempts on isolated-word small vocabulary speech recognition, mostly using dynamic programming techniques. For example, dynamic time warping was used to develop a small vocabulary isolated word recognizer (Lipeika, 2002).

Lately, a hidden Markov model based medium vocabulary speaker independent isolated-word recognizer was reported to perform at 20% word error rate (Raskinis, 2003).

There have been no results for Estonian large vocabulary continuous speech recognition yet. There are two main reasons for this. First, very large speech and text corpora are not available yet, although there is ongoing active work in that area. Second, the Estonian language poses some specific difficulties that make building a large vocabulary speech recognition system complicated. Estonian is an agglutinative language, thus it's words are heavily inflected depending on their syntactic role. This makes the number of distinctive words in the language very large. Therefore, a high out-of-vocabulary (OOV) rate is expected when words are used as recognition units in composing a statistical language model (Weibel, 2000). Also, the word order in Estonian is much more free than in non-agglutinative languages like English, which also complicates building of a good language model.

The final objective of our work is to build a large vocabulary continuous speech recognition system for Estonian. To isolate the problem of composing a language model for large vocabulary speech recognition from the problem of acoustic modelling, we decided to start with using artificial language models of limited size. This approach still enables to build applications for real-world use but doesn't require any long work in the area of language modelling. Also, the use of fixed language models lets to focus on the problems and methods of phoneme and word modelling which are the foundations of a speech recognizer.

In this paper we present continuous speech recognition experiments based on one small and one medium-scale limited language model. For all experiments, we use the Estonian part of the BABEL multi-language speech database which is introduced in the first section of the paper. The second section describes the process of extracting acoustic features from human speech, some important phonetic properties of Estonian and the process of training acoustic models. In the fourth section, we present results of recognition experiments using an Estonian number-recognizer and a medium-sized word-loop grammar. The paper concludes with an analysis of the results and the description of future work.

2. Resources

2.1. *Training Data*

A large speech corpus is needed for training speaker independent acoustic models. The only systematically composed speech database available for Estonian is the Estonian Phonetic Database that is part of the BABEL multi-language database (Eek, 1999). The corpus is partitioned into 3 sets: many talker set (30 male and 30 female speakers), few talker set (4 male and 4 female speakers) and a very few talker set (a male and a female speaker). Speech recordings are performed in an anechoic room and digitized at 20 kHz and 16 bit. For each speaker, there are one or more recorded text passages, a set of isolated utterances, and a set of isolated read numbers. The texts that were read were selected so that all main phonologically relevant oppositions would be revealed in

the corpus. All recordings come with sentence level transcriptions in both orthographic as well as SAMPA phonemic formats. In addition to the mentioned recordings, there are also isolated CVC construction recordings in the database, which were however not used for training the recognizer.

The text passage and sentence recordings in the many talker set were used for training the acoustic models. This resulted in 1230 audio files, approximately 3:45 of duration. The recordings of read numbers in the few talker set (400 utterances, 23 minutes of speech) were used for evaluating the performance of the number recognition system. The isolated sentence recordings in the few talker set, approximately 16 minutes of speech, were used for evaluating the limited vocabulary speech recognizer.

2.2. *Hardware and Software*

The recognition system is based on the Hidden Markov Model Toolkit (HTK) (Young, 2003), version 3.1 (and later 3.2). Various scripts in the J programming language (Iverson, 2001), as well as Perl and shell scripts were used to prepare and convert training data, and to automate the training and evaluation process.

The training and recognition experiments were run on a standard PC with an AMD 1.0 GHz Athlon processor, 256 MB RAM, running Linux Mandrake.

3. Properties of Estonian

3.1. *Selection of Phonological Units*

Estonian phonemic inventory is quite small. There are nine vowels, each of which corresponds to a single grapheme. They can be grouped with respect to the tongue position, tongue height and the roundedness of the lips in their articulation process. All those features are important for building accurate context-sensitive acoustic models. All vowels are shown in Table 1, together with their symbols according to IPA and SAMPA represen-

Table 1

Estonian vowel inventory

IPA	*SAMPA*	*Estonian phonological transcription*	*Example word*	*Tongue position, tongue height, lips*
ɑ	A	/a/	*sada*	back, low, unrounded
e	e	/e/	*keda*	front, medium-high, unrounded
i	i	/i/	*kilu*	front, high, unrounded
o	o	/o/	*pori*	back, medium-high, rounded
u	u	/u/	*kuri*	back, high, rounded
ɤ, ɯ, ə	7	/õ/	*kõma*	back, high, unrounded
æ	{	/ä/	*käru*	front, low, unrounded
ø	2	/ö/	*löma*	front, medium-high, rounded
y	y	/ü/	*mürin*	front, high, rounded

tation. There is no single appropriate character for the Estonian /õ/ in the IPA inventory, although [γ] seems to be the most common (Eek, 1999).

There are 6 plosives (/p/, /b/, /k/, /g/, /t/, /d/) , 2 nasals (/m/, /n/), a thrill (/r/), 3 fricatives (/v/, /s/, /h/), a lateral (/l/) and an approximant (/j/) in native Estonian. In addition, the phonemes /f/ and /š/ occur in foreign words. Phonemes /t/, /s/, /n/ and /l/ occur as palatilised in many words. In the given recognizer, unpalatalised and palatilised variants of those phonemes are modelled as single units. The rationale behind this is that those phoneme variants correspond to the same character in orthographic form, and in addition, this simplifies building the pronouncation dictionary for the recognizer. However, modelling those phonemes as separate units should be considered in the future, as the (un-)palatalisation often determines the semantics of the word, and this information is important for language modelling.

The consonant phonemes together with their articulation characteristics are listed in Table 2.

In addition to the listed phonemes, there are secondary appearances of some of the consonants: e.g., /n/ is often realized as a velar [η] before /g/ and /k/. Sonorants /m/, /n/, /l/, /r/ and /v/ are sometimes idiosyncratically devoiced when preceded by /h/ at the end of one-syllable words (Eek, 1999). Those secondary realizations are not separately

Table 2

Estonian consonant inventory

IPA	*SAMPA*	*Estonian phonological transcription*	*Example word*	*Type of articulation, voiced/voiceless, place of articulation*
b̥	p	/b/	*taba*	unintensive plosive, voiceless/half-voiced, bilabial
p	pp	/p/	*tapa*	plosive, voiceless, bilabial
d̥, d̥^{j}	t, t’	/d/	*padu, padi*	(palatilized) unintensive plosive, voiceless/half-voiced, denti-alveolar
t, t^{j}	tt, t’t	/t/	*patu, pati*	(palatilized) plosive, voiceless, denti-alveolar
g̥	k	/g/	*kagu*	unintensive plosive, voicelesshalf-voiced, palato-velar
k	kk	/k/	*kaku*	plosive, voiceless, palato-velar
f	f	/f/	*foori*	fricative, voiceless, labiodental
v	v	/v/	*kava*	fricative, voiced, labiodental
s, s^{j}	s, s’	/s/	*mäsu, kasi*	(palatilized) fricative, voiceless, alveolar
ʃ	S	/š/	*šefi, looži*	fricative, voiceless, postalveolar (usually labialized)
h	h	/h/	*sahin*	fricative, voiceless, glottal-oral; short degree-voiced lenis; long degree-voiceless geminate
m	m	/m/	*samu*	nasal, voiced, bilabial
n, n^{j}	n, n’	/n/	*kanu, pani*	(palatilized) nasal, voiced, alveolar
l, l^{j}	l, l’	/l/	*kalas, pali*	(palatilized) lateral, voiced, alveolar-postalveolar
r	r	/r/	*nari*	thrill, voiced, alveolar
j	j	/j/	*maja*	approximant, palatal

modelled in the given recognizer. As those phoneme variants occur in certain contexts only, they should get automatically modelled when context-sensitive phoneme models of reasonable complexity are trained.

3.2. *Quantity Degrees*

All vowels and consonants, with some exceptions, occur in short, long and overlong quantity degrees. Quantity degrees, together with other prosodic features such as stress and tone, play a very important and multifunctional role in the Estonian language.

Quantity degrees are directly responsible for determining the semantics of the word. For example, in the word *kalu/kaalu*, the vowel /a/ can be realized as a short, long or an overlong phoneme, each of which would result in a different meaning (/kalu/ 'fish', part. pl.; /kaalu/ 'weight', gen. sg.; /kaa:lu/ 'weight', part. sg.).

Different quantity degree representation of the same phoneme type are not modelled as separate phonological units rather than one or a sequence of two instances of the same phoneme. The only difference are plosives /k/, /p/, /t/: short unintensive realizations of those phonemes are modelled as separate units /g/, /b/, /d/. However, those units do not directly correspond to voiced variants of those plosives, as in many other languages. In our case, those phonemes are just short and unintensive but still voiceless (sometimes also referred to as half-voiced) representations of the given plosives.

A phoneme occurrence in second and third quantity degree are both modelled as a sequence of two instances of the same phoneme. There are two pragmatic reasons behind this: second and third quantity degrees are usually not differentiated in orthographic form, with the exception of plosives (but the approach to plosives is described in the previous paragraph). Secondly, it turns out that long and overlong quantity degrees cannot be identified only by observing the characteristics of the phoneme realization; quantity degree is rather a property of the whole foot, and is determined by many prosodic variables of the neighbouring syllables (Eek, 1997). Thus, they cannot be modelled with phone-level HMMs, a more complex modelling of the cross-syllable prosodic features of speech is required. This is a topic for our future work.

4. System Description

4.1. *Feature Extraction*

Raw speech waveforms were parameterized into sequences of feature vectors using Mel Frequency Cepstral Coefficients (MFCCs). One feature vector was computed in every 10 ms, using a window size of 25 ms. Hamming window was used in calculating the FFT, and a first order preemphasis coefficient of 0.97 was applied. For each frame, 12 MFCC coefficients were computed from the output of 26 mel scale filterbanks. The log of the signal energy in the observed window was used as the 13th vector component.

In addition to the static components, also first and second order regression coefficients (also known as delta and acceleration coefficients) were computed by comparing the

features of the observed window with those of its left and right neighbour. As a result, a 39 component feature vector corresponds to each frame.

4.2. *Acoustic Models*

The system's basic units of recognition are phonemes which are modelled by hidden Markov models (HMMs). 22 phonemes, a silence and a possible pause between words were modelled. Phoneme models have three emitting states and a left-to-right topology. Silence model has a backward skip from the last emitting state to the first in order to better absorb various noises in the data (Young, 2003) . Short pause has one emitting state that is shared by the middle state of the silence model, and a skip transition over the only emitting state, which makes it a so-called *tee-model* (Young, 2003).

4.3. *Training*

A flat start training strategy was used to initialize the monophone models: global mean and variance were computed from the training data and the parameters of all models were set equal to those components. The HTK implementation of the Forward-Backward algorithm was then used to form the parameter estimates of all models. Several iterations of the algorithm were applied to get a final set of monophone models.

From the trained monophone set, models for all cross-word triphones that were present in training data were created. The triphones were initialized by cloning the corresponding monophone models and re-estimated in multiple iterations.

The next step in the training process was to tie states within triphone sets in order to share training data and thus be able to make models more robust. Acoustic decision trees based on phoneme properties described in Section 3 were used to find candidates for triphone state clustering (Young, 1994). This also allowed to synthesize models for all triphones that do not occur in the training data (unseen triphones). This resulted in a total of 9092 unique HMM states shared by all different models.

The last step in the model building process was the conversion from single Gaussian HMMs to multiple mixture component HMMs. The number of mixture components were increased iteratively: after each mixture splitting, a couple of re-estimation iterations were done. This was repeated until each state consisted of six Gaussian components.

The same set of acoustic models was used for both evaluation tasks. The models can be used for recognizing any Estonian speech, given a word pronouncation dictionary and a language model.

4.4. *Language Models*

For number recognition, a grammar in Backus–Naur Form (EBNF) was specified. The grammar was composed so that all Estonian numbers from 0 to 1 999 999 would be accepted with equal probability. The grammar was compiled into a word network that could subsequently be used by the HTK recognizer.

For limited vocabulary recognition experiment, a list of all unique word forms was extracted from the transcriptions for speech data that was going to be used for recognition experiments. The resulting list of 1887 words was used to build a word-loop language model. Thus, we got a language model that accepted all word sequences with an equal probability, given that the words occurred in transcriptions.

5. Experimental Results

5.1. *Number Recognition*

When evaluating the performance of the number recognizer, the measure of correctly recognized numbers was used, as opposed to the word error rate that is more often used in speech recognition. This was done because in number recognition it is important to recognize all digits (or words in this case) correctly – even a one digit error is not acceptable for most real tasks.

Also, recognition speed measured as recognition time divided by the duration of the test utterances (giving the realtime performance of the system) was used in evaluating the system.

Recognizer performance was measured with acoustic models of increasing complexity. Table 3 lists number recognition accuracy and realtime performance with clustered triphone models having one to six Gaussian mixtures. As can be seen, an accuracy of 97.0% was measured with most complex models. The time needed for processing was only 5% of the duration of the test utterances with even the most complex models.

5.2. *Recognition of Limited-Vocabulary Speech*

The recognition quality was measured by comparing output of recognizer with the reference transcription. As the common approach in speech recognition is, recognized and reference word sequences are matched using dynamic programming to obtain an optimal string match. This alignment provides a number of word substitution errors (S), deletion

Table 3
Number recognition performance

Mixtures	*Accuracy (%)*	*Realtime*
1	88.3	0.01
2	93.5	0.02
3	93.8	0.03
4	96.8	0.04
5	96.5	0.04
6	97.0	0.05

errors (*D*) and insertion errors (*I*) for each sentence. There are two commonly used figures that can be calculated from these numbers: the percentage of correctly recognized words is calculated as

$$C = \frac{N - D - S}{N} \times 100\%, \tag{1}$$

where *N* is the number of words. Another measure – recognition accuracy – also considers insertion errors:

$$A = \frac{N - D - S - I}{N} \times 100\%. \tag{2}$$

As the realtime performance of the recognizer was considered important, the recognition network pruning feature of the HTK decoder was employed. The pruning is implemented by deactivating branches of the search network whose log probability fall more than a *beam-width* below the best (Young, 2003). A beam-width of 180 was experimentally found to give good recognition accuracy while still resulting in real-time performance.

The limited vocabulary recognizer achieved real-time word recognition correctness of 82.9%, and an accuracy of 81.9%. Some tests were also done to evaluate the non-realtime performance of the recognizer. After increasing the pruning beam width to 700, the recognition correctness rose to 90.6%. However, this was achieved with increased CPU time which was 1700% of the duration of test utterances. Table 4 shows recognition performance with varying model complexity.

When analyzing the recognition accuracy results more closely, an interesting detail was found. Namely, the speech of male speakers was recognized significantly better than that of female speakers, as can be see on Fig. 1. The recognition of male speech had an overall percentage of correctly recognized words of 86.9, while the same value for female speakers was only 79.0. This is so despite the fact that male and female speech was quite equally represented in training data. Such observation can be grounded on the statistics that show significantly more acoustic variability in female speech than in male

Table 4

Word recognition performance

System description	*Correct (%)*	*Accuracy (%)*	*Realtime*
1 mixture clustered triphones (CT)	74.1	71.2	0.95
2 mixture CT	79.2	76.6	0.87
3 mixture CT	80.7	79.2	0.85
4 mixture CT	81.4	79.9	0.85
5 mixture CT	82.6	81.1	0.85
6 mixture CT	82.9	81.9	0.86
6 mixture CT, loose pruning	90.6	89.1	16

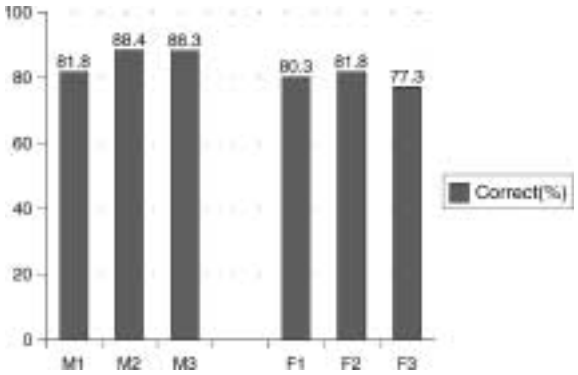

Fig. 1. Recognition accuracy of male and female speech.

speech (Syrdal, 1996). This leads to the idea that relatively more female speech should be available for robust acoustic model training. Another technique that could be used to fight this effect is vocal tract normalization (VTN) that is found to efficiently compensate for the acoustic variability due to different vocal tract lengths of various speakers (Lee, 1996).

6. Discussion

The number recognition results are satisfactory but not groundbreaking. Higher number recognition results have been reported for many languages and with many different modelling approaches. However, we didn't train the models specifically for number recognition, but used common phoneme models that were concatenated to part-of-number models. Using digits and other parts of numbers as recognition units could have given a much higher recognition rate, but this wasn't in the scope of our work.

The fixed medium-sized vocabulary recognition results should be analyzed with respect to the used language model and its perplexity. Perplexity measures the cross-entropy of the language models and can be regarded to show how many different most probable words can follow any given word (Young, 2003). A language model with perplexity X has roughly the same difficulty as a word-loop language model of X words (Huang, 2001). Thus, our word loop system performance can be compared with other systems where a language model perplexity of roughly 1887 was used. It turns out that such perplexity is extremely high: typical perplexities of n-gram models for English range from about 50 to at most 1000. Dialogue systems, such as a 2000-word conversational Air Travel Information System, have perplexities typically less than 20 (Huang, 2001). A high perplexity also causes much increased computing time because many word candidates must be considered at each frame. Thus, it causes a tighter pruning beam and reduced recognition quality if realtime performance is considered important. Given such background, our realtime recognition correctness of 82.9% is relatively good. However, it must be noted that perplexity cannot take into account acoustic confusability of words, and thus such direct comparison is not quite accurate.

7. Conclusions and Future Work

We built a prototype of a limited vocabulary Estonian speech recognition system. Context-dependent triphones modelled with hidden Markov models were used as basic units of recognition. The recognition performance was evaluated on two tasks. Number recognition system reached a recognition accuracy of 97%. A medium vocabulary system using a vocabulary of 1887 different word phrases and a word loop grammar as language model recognized 82.9% of words correctly, when operating in real-time. After giving up the the real-time requirement, a word correctness of 90.6% was achieved.

Future works will be focused on language modelling in order to build an Estonian large vocabulary speech recognition system. Our aim is to develop methods to overcome the problem of recognizing a strongly agglutinative language, especially Estonian.

Future plans also include studying the techniques to model the three different word foot quantity degrees and other accentual and prosodic factors that play an important role in Estonian language.

References

Eek, A., and E. Meister (1997). Simple perception experiments on estonian word prosody: foot structure vs. segmental quantity. In *Estonian Prosody*: *Papers from a Symposium*, Tallinn. pp. 71–99.

Eek, A., and E. Meister (1999). Estonian speech in the BABEL multi-language database: Phonetic-phonological problems revealed in the text corpus. In *Proceedings of LP'98*, Vol. II. pp. 529–546.

Huang, X., A. Acero, H.-W. Hon and R. Reddy (2001). *Spoken Language Processing*: *A Guide to Theory, Algorithm and System Development*, 1st edition. Prentice Hall PTR.

Iverson, E. (2001). *J Primer*. Jsoftware Inc.

Lee, L., and R. Rose (1996). Speaker normalization using efficient frequency warping procedures. In *Proceedings of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Atlanta, USA. pp. 353–356.

Lehiste, I. (1966). Consonant quantity and phonological units in Estonian. *Indiana University Publications, Uralic and Altaic Series*, **65**.

Lipeika, A., J. Lipeikienė and L. Telksnys (2002). Development of isolated word speech recognition system. *Informatica*, **13**(1), 37–46.

Meister, E., B. Lobanov, R. Vahisalu, T. Levkovskaya, V. Kisialou, P. Tatter and J. Lasn (2001). Spoken dialogue system for mobile parking. In *Proceedings of the International Workshop SPEECH and COMPUTER (SPECOM 2001)*, Moscow, Russia, October 2001. pp. 123–126.

Meister, E. (2001). Towards speech recognition in Estonian. 21. Fonetiikan Päivät, Turku 4.-5.1.2001. In S. Ojala and J. Tuomainen (Eds.), *Publications of the Department of Finnish and General Linguistics of the University of Turku*. pp. 59–70.

Raskinis, G., and D. Raskiniene (2003). Building medium-vocabulary isolated-word Lithuanian HMM speech recognition system. *Informatica*, **14**(1), 75–84.

Syrdal, A.K. (1996). Acoustic variability in spontaneous conversational speech of American English talkers. In *Proc. ICSLP '96*, Philadelphia, PA, Vol. 1. pp. 438–441.

Waibel, A., P. Geutner, L.M. Tomokiyo, T. Schultz, and M. Woszczyna (2000). Multilinguality in speech and spoken language systems. *Proceedings of the IEEE*, **88**(8), 1297–1313.

Young, S., J. Odell and P. Woodland (1994). Tree-based state tying for high accuracy acoustic modelling. In *ARPA Workshop on Human Language Technology*. pp. 307–312.

Young, S., G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland (2003). The HTK Book (for HTK Version 3.2).
`http://htk.eng.cam.ac.uk`.

T. Alumäe was awarded the Master of Technical Sciences Degree at Tallinn Technical University in 2002. At present he is a researcher in the Laboratory of Phonetics and Speech Technology of the Institute of Cybernetics at Tallinn Techical University. Scientific interests include: speech and language processing, signal processing, artificial intelligence.

L. Võhandu is a professor emeriticus at Tallinn Technical University. As a scientist of mathematics and informatics, he has played an important role in the developement of technical sciences in Estonia. Besides countless scientific publications, he is an author and co-author of several books and numerous essays and articles. His scientific interests include: data processing and data mining, random processes, statistics, self-organization and multimedia.

Riboto žodyno ištisinės estų kalbos atpažinimo sistema, naudojanti paslėptus Markovo modelius

Tanel ALUMÄE, Leo VÕHANDU

Straipsnyje pateikta riboto žodyno nepriklausanti nuo kalbėtojo ištisinės estų kalbos atpažinimo sistema, besiremianti paslėptais Markovo modeliais. Sistema yra apmokyta naudojant anotuotą 60 kalbėtojų estų kalbos duomenų bazę, kurios balso įrašų trukmė yra apytikriai 4 valandos. Žodžiai yra modeliuojami naudojant klasterizuotus trigarsius su keliomis Gauso mišinių komponentėmis. Sistema yra testuota atpažįstant skaičius ir vidutinio dydžio žodyno ištisinę kalbą. Sistemos darbingumas tirtas naudojant didėjančio sudėtingumo akustinius modelius. Atpažįstant skaičius buvo pasiektas 97% atpažinimo tikslumas. Kai atpažįstama realiame laike, vidutinio dydžio žodyno atpažinimo sistema atpažįsta 82.9% žodžių. Kai atsisakoma atpažinimo realiame laike reikalavimų, atpažinimo tikslumas padidėja iki 90.6%.