Real-time spontaneous Ukrainian speech recognition system based on word acoustic composite models

Valentyna Robeiko, Mykola Sazhok

Speech Science and Technology Department, International Research and Training Center of Information Technologies and Systems; CyberMova; Kyiv, Ukraine
{ valya.robeiko, sazhok } @gmail.com

Abstract

This paper describes implementation of methods and algorithms for the automatic speech recognition based on word composition proceeding from acoustic phoneme models. Such a design of the speech-to-text decoder is conventional and most productive for Western languages. The aim is to explore this approach applied to the Ukrainian language that is highly inflective with relatively free word order. We use data-driven methods to estimate parameters for both acoustic and linguistic components of the mathematical model. The grapheme-to-phoneme conversion procedure takes into account word stress issue and spontaneous continuous speech features. The basic speech-to-text system is able to operate a 100k vocabulary in real-time. The prospective of dictionary and domain extension, parameter estimation improvement and ergonomic issues are discussed.

Index items: Speech recognition, spontaneous continuous speech, generative model, real-time.

1. Introduction

Speech recognition systems gradually possess a place of mediator between a human and a computer overriding the habitual means of information and knowledge collecting and exchange. For English, next to the dictation software on the PC, a number of network systems have been introduced that serve the input for voice searches or allow you to dictate an e-mail [1]. These systems demonstrate quite reasonable efficiency even with a significant delay paid to the network for speech signal delivering to cloud-services. Apparently, these systems (a) operate with large lexicon and (b) provide the real-time computing.

Analysis of proprietary patents and publications of leading research centers shows that the most referred HMM-based recognition scheme considers generation of composite speech patterns for words or phrases consisting of the phoneme acoustic models at the stage of acoustic decoding [2], [3]. Simultaneously, proceeding from the linguistic model, the likelihood of hypothetically recognized partial word sequences is estimated by the preceding one or more words and merged to the integral likelihood.

Specific features of Slavonic languages are high inflectiveness and relatively free word order. This leads to rapid growth of the recognition vocabulary (8-10 times larger for same English lexicon) and weakening of the language model prediction force. That is why the applicability of conventional methods and algorithms to Slavonic languages looks rather unpromising. That is the reason of search for new recognition schemes, particularly considering word composition by the acoustic decoding output [4].

Contemporary speech-to-text systems can recognize isolated words and red speech like news reading with about 5% word error rate [1], [3]. Still conversational or spontaneous speech recognition is far less reliable. Spontaneous speech recognition in realistic conditions of communication (e.g. background noise) is an extremely actual problem, which solution would significantly extend the speech recognition system application.

However new approach development is indispensable, still it remains uncovered the potential of the recognition scheme having been developed for decades [2], [3]. The open question is limits of the lexicon used in the speech-to-text system based on the conventional recognition scheme provided that the system shows real-time performance on a computational platform available for an ordinary user. Therefore we aimed to build a real-time system that can be exploited on a contemporary personal computer for speech-to-text conversion like a dictation machine.

In the next Section we describe structures and means for speech recognition data and knowledge base, the choice of speech and language data is justified, grapheme-to-phoneme conversion and spontaneous speech features are discussed. In Section 3 an acting speech-to-text system is described, its features and possible applications are covered. In Conclusion we consider most actual improvements, the current state and of research is discussed and future plans are revealed.

2. Speech signal generative model parameters and their estimation

The input speech signal from is converted to a sequence of fixed size acoustic vectors or features $\mathbf{Y}_{1:T} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T)$ by the pre-processor. In other words we translate a waveform to the feature space. Then the decoder attempts to find the word sequence $\mathbf{w}_{1:L} = (w_1, w_2, ..., w_L)$ that most likely corresponds to the observed $\mathbf{Y}$. In other words the decoder has to find:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}} P(\mathbf{w} \mid \mathbf{Y}). \quad (1)$$

Despite complexity, some models, which are based on discriminative approach, try modeling this expression directly [5]. However, a more productive way used in generative model is to apply Bayes' Rule to (1) and consider the equivalent problem of finding:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}} p(\mathbf{Y} \mid \mathbf{w}) P(\mathbf{w}). \quad (2)$$

Here the likelihood $p(\mathbf{Y}\,|\,\mathbf{w})$ is an acoustical component and the probability $P(\mathbf{w})$ is a linguistic component of the speech recognition generative model.

The acoustic component is determined by an **acoustic model** (AM). Each spoken word w is decomposed into a sequence of L_w basic sounds called basic phones. This sequence is a pronunciation of the word or its phoneme transcription $\mathbf{q}_{1:K_w}^{(w)} = (q_1, q_2, ..., q_{K_w})$. Developing a speech technology we must consider individual and situational speaker peculiarities, word pronunciation in the speech flow, which causes introduction of the multi-decision into grapheme to phoneme conversion. To allow for the possibility of multiple pronunciations, the $p(\mathbf{Y}\,|\,\mathbf{w})$ can be computed over multiple phoneme transcriptions:

$$p(\mathbf{Y}\,|\,\mathbf{w}) = \sum_{\mathbf{Q}} p(\mathbf{Y}\,|\,\mathbf{Q}) P(\mathbf{Q}\,|\,\mathbf{w}), \qquad (3)$$

where summation is over all valid pronunciation sequences for **w** and **Q** is a particular sequence of pronunciations,

$$P(\mathbf{Q}\,|\,\mathbf{w}) = \prod_{l=1}^{L} P(\mathbf{q}^{(w_l)}\,|\,w_l), \qquad (4)$$

where $\mathbf{q}^{(w_l)}$ is a valid pronunciation for w_l. In practice, the expression (3) is computed using a maximum instead of summation and minimizing a number of alternative pronunciations in (4) we reach essential reduction of composite models if alternative pronunciations are introduced conservatively.

Conventionally, an acoustical essence of the phoneme q is expressed in form of a generative model as shown in Fig. 1a where $\{a_{ij}\}$ are statistical parameters of transition between states, $\{b_j(\,)\}$ are probabilistic distributions in feature vector space for the emitting states. In fact, the distributions approximate regions the speech signal corresponding to a q moves through in the feature space. Such is a representation of a basic HMM. Technically, transition from an emitting state to any other state directly connected to the emitting state takes a time sample. Matrix $\{a_{ij}\}$ depends on an HMM topology and is a stochastic matrix that form a Markov chain.

A valid sequence of states

$$\Theta_{1:T} = (\theta_1, \theta_2, \ldots, \theta_T), \qquad (5)$$

that generates a model signal is an acoustic transcription of the observed signal. In accordance to generative model, these states have conditional relations between themselves and observed signal samples. На рис. 1a these dependencies are represented in form of Dynamic Bayes Network (DBN) [3]. In the DBN notation used here, squares denote discrete variables; circles continuous variables; shading indicates an observed variable; and no shading an unobserved variable. The lack of an arc between variables shows conditional independence.

Such a form of presentation is convenient to illustrate extensions of the basic generative model structure. It is simple to introduce new unobserved parameters and dependences e.g. between neighboring samples of the observed signal. Moreover, DBN is useful for explanation the speech recognition discriminative models.

To approximate phoneme observation regions via output probabilities $\{b_j(\,)\}$, instead a single multivariate Gaussian $G(\mathbf{y};\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the Gaussian mixture model is introduced:

$$b_j(\mathbf{y}) = \sum_{m=1}^{M} c_{jm} G(\mathbf{y};\boldsymbol{\mu}^{(jm)}, \boldsymbol{\Sigma}^{(jm)}) \qquad (6)$$

where c_{jm} is the prior probability of staying in component m of the state j. These priors satisfy the standard constraints for a valid probability mass function: $c_{jm} \geq 0$ i $\sum_{m=1}^{M} c_{jm} = 1$.

A mixture of Gaussians is a highly flexible distribution able to model asymmetric and multi-modal distributed data. This allows for better representation of the speech signal diversity on acoustic level.

Another important issue is to provide the diagonality for each co-variation matrix $\boldsymbol{\Sigma}^{(jm)}$ in order to optimize computational cost and justify the acoustic training data capability. The digital cosine transformation applied to spectral coefficients in cepstral analysis is sufficient in most cases. Thus we approximate phoneme observation regions uniting ellipsoids stretched alongside the coordinate axes of feature space.

In Fig. 2 shows a feature space trajectory projection to 2D space for the spoken word sample *oca* ("a wasp" pronounced as "*oh s ah1*"). The observed sequence of acoustic vectors $\mathbf{y}_{t=1:72}$ bypasses regions of matching to specific phonemes: # (phoneme-pause), *o*, *c*, *A* ("ah" stressed) and #. The region for the pause # is approximated with an ellipsoid corresponded to the single Gaussian of the only state #1 for the pause. We admit that the probability of

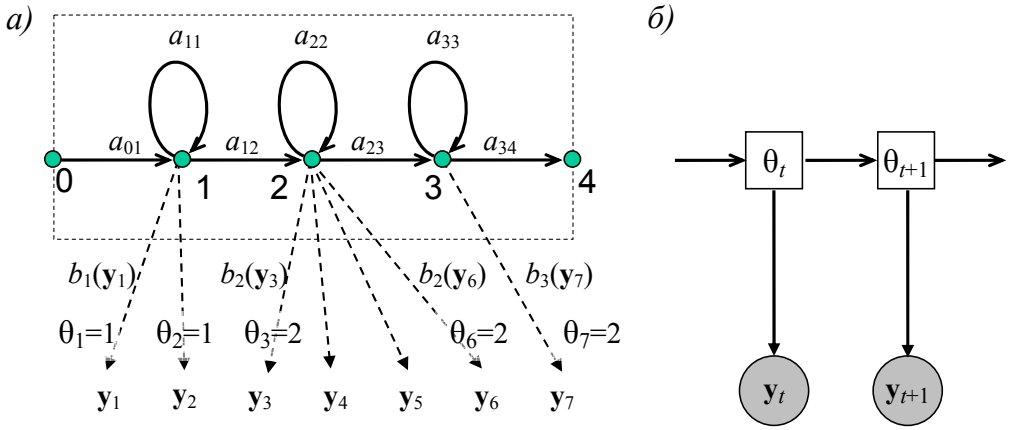

Figure 1. Basic generative model HMM for the phoneme a) as a wrapped dynamic programming graph, and b) in terms of Dynamic Bayes Network

the acoustic vector observation given a Gaussian is greater than 0.1 inside the relevant ellipsoid. Models for phonemes *o* and *A* contain three states: *o*1, *o*2, *o*3 and *A*1, *A*2, *A*3, each of which has a distribution modeled with a two-component Guassian mixture. Gaussians corresponding to the same phoneme state have identical stroke. The model of the phoneme *c* consists of three states as well; however, each of states is modeled just with a single Gaussian.

At the stage of decoding the dynamic programming procedure searches among valid acoustic transcriptions (5) the one that approximates the signal trajectory in a best way. The fines approximation for the signal trajectory shown in Fig. 2 is the acoustic transcription specified with values: $\theta_{1:10} = \#1$, $\theta_{11:16} = o1$, $\theta_{17:24} = o2$, $\theta_{25:27} = o3$, $\theta_{28:31} = c1$, $\theta_{32:39} = c2$, $\theta_{40:42} = c3$, $\theta_{43:48} = A1$, $\theta_{49:59} = A2$, $\theta_{60:66} = A3$, та $\theta_{67:72} = \#1$.

The black circle radius at the point of a Gaussian expectation corresponds to integral value for the staying in same state *j* probability a_{jj} and the output probability $b_j(\,)$ according to (6). An integer span label indicates time samples having the finest approximation the Gaussian provide. Gaussians having no assigned such a label still contribute to the output probability but this amount is the least. A square marker indicates a particular point at the trajectory. The closest Gaussian center is connected to the square maker with a solid line.

Acoustic model parameters $\{a_{ij}\}$ and $\{b_j(\,)\}$ are estimated from a corpus of training utterances iteratively. Firstly, a single Gaussian model is introduced. Then gradually the number of Gaussians is being increased by splitting those having the biggest norm of co-variation matrix. The maximal number of Gaussians is estimated proceeding from the condition of 50 phoneme samples per Gaussian.

A linguistic component of the model (2) consists in estimating the probability:

$$P(\mathbf{w}) = \prod_{k=1}^{K} P(w_k \mid w_{k-1},...,w_1). \qquad (7)$$

However, initially no restrictions are introduced on the number of preceding words, in practice it is limited till $N-1$ and hence the *N*-gram Language Model (LM) can be formulated:

$$P(\mathbf{w}) = \prod_{k=1}^{K} P(w_k \mid w_{k-1}, w_{k-2},..., w_{k-N+1}), \qquad (8)$$

where *N* is chosen normally between 2 and 4. Probabilities of *N*-grams are estimated from training texts by counting *N*-gram occurrences to form maximal likelihood parameter estimates. For example, if $C(w_{k-N+1},...,w_{k-1},w_k)$ represents the frequency of *N*-grams $(w_{k-N+1},...,w_{k-1},w_k)$, then

$$P(w_k \mid w_{k-1}, w_{k-2},..., w_{k-N+1}) \approx \frac{C(w_{k-N+1},...,w_{k-1},w_k)}{C(w_{k-N+1},...,w_{k-1})}. \qquad (9)$$

The major problem with this formulation of LM is probability estimation for *N*-grams that have insufficient statistics. Then such estimations are completed proceeding from (*N*–1)-grams [3]. The other problem is presence of words excluded from the recognition vocabulary (OOV words). An acceptable solution is introduction an *unknown word* category that replaces all OOV words. Furthermore, large size of LM can be the obstacle for speech recognition system practical implementation.

3. Speech recognition system implementation and trial operation

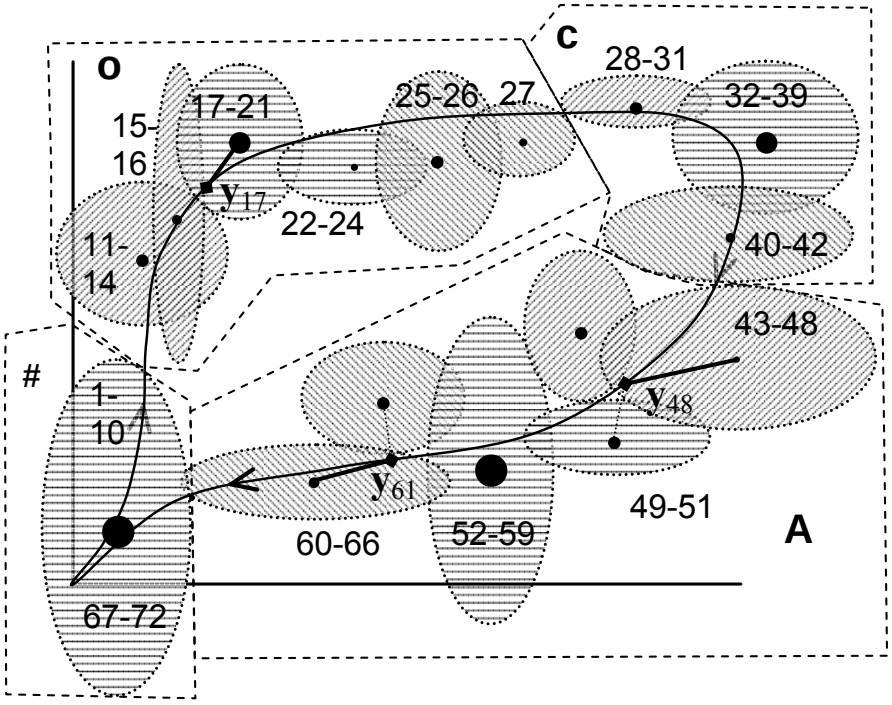

Рисунок 2. A feature space trajectory projection to 2D space for a spoken Ukrainian word *ocá* sample, which is pronounced as # *o c A* # (in English, the word "a wasp", which is pronounced as "*oh s ah1*")

The basic speech-to-text conversion system structure is shown in Fig. 3. The real-time modules implement a proper decoder and modules for the mathematic model parameter estimation are run in postponed mode. To develop the speech recognition system we used own program modules and several toolkits available on Internet like: *HTK*, *HTS*, *Julius*, *MITLM*, *CMU LM* [9], [10], [11].

Real time component takes the *Input speech signal* from an available source (microphone or file system). Passing through *Voice activity detector* the signal is segmented into portions where speech presence is assumed. To make such an assumption simple features based on amplitude and zero-crossing count are used. *Pre-processor* extracts acoustic features from the speech signal segment. Features are based on mel-frequency cepstral coefficients with subtracted mean and accomplished with energy and dynamic components (delta and delta-delta coefficients). Finally, a feature vector dimension is equal to 39. *Decoder* compares an input segment with model signal hypotheses generated in accordance to (2)–(8), using a conservative strategy of non-perspective hypotheses rejecting [10]. Both acoustic and linguistic components are exploited to accomplish the decoding process. A sequence of words that generates a model signal that is most similar to the input signal is assigned to *Recognition response*.

System vocabulary of the system consists of frequency dictionary extracted from the text corpus and supplementary vocabularies (speech corpus training sample, social and local dialects, proper names, abbreviations etc.). In spite of English, the Ukrainian alphabet includes both stressed and non-stressed vowels and palatalized and non-palatalized counterparts for most consonants. Word stress position is retrieved from the ULIF dictionary [13], for supplementary vocabularies stress is pointed either by an expert or automatically predicted [8]. Most frequent one- or two-syllable words are presented by both stressed and non-stressed alternatives.

Grapheme-to-phoneme module converts words accomplished with stress information to phoneme sequences used to form word composite acoustic models for both decoder and AM parameter estimation. This module implements a multi-decision symbol conversion technique [7], which allows for modeling the distinctions proper to a specific language and is based on learning the regularities of relation between orthographic and phonemic symbols. The expert formulated about 40 local rules of grapheme-to-phoneme conversion partially modeling the individual speaker peculiarities and co-articulation and reduction of sounds in speech flow. On average each word produces 1,2 transcriptions. The multi-level extension to the conversion technique also allows for expressing numbers and symbolic characters in Ukrainian words.

Text corpus for the language model is based on a hypertext material downloaded from several websites containing samples of news and publicity (60%), literature (8%), encyclopedic articles (24%) and legal and forensic domain (8%). To be noted that the data downloaded from news websites contains numerous user comments and reviews, which mostly are text samples of spontaneous speech.

Text filter, used for text corpus processing, allows for converting numbers symbolic characters to relevant words, removing improper text segments, paragraph repetitions and sentences containing a particular percentage of words missing in the ULIF dictionary. Total size of *Text corpus* is 2 GB that includes 17,5 million sentences that is a list of words containing above 250 million items and forming a vocabulary of two million words.

Processed text enters the toolkit that forms an *N*-gram based *Language model* given a recognition vocabulary. Initially, sentences containing above the specified percentage OOV words are removed and such words in other sentences are marked as *unknown*. The maximal order of the created model is 3. For the recognition vocabulary of 100 000 words total 88,5 million distinct 3-grams are detected, unknown words occupy 2,5% of the 1,2 GB sized data. To model spontaneous speech characteristics a class of *transparent* words is introduced. It contains non-lexical items like pause fillers and emotion and attitude expressions.

The real-time component is used to build a basic speech-to-text conversion systems used in experimental research. Graphical user interface integrated with the basic system allows for demonstrating continuous speech recognition for wide domain in real time using a contemporary notebook (рис. 3) [14].

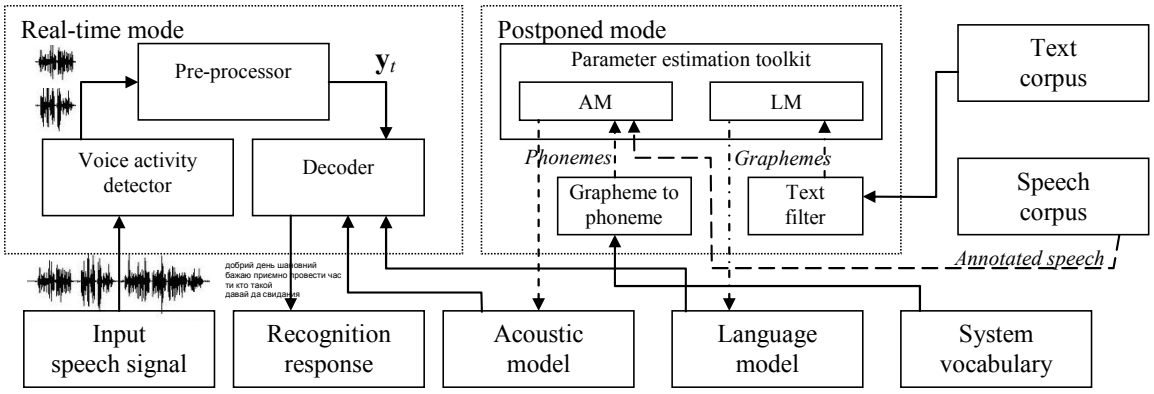

Figure 3. General structure for the basic speech-to-text system

The developed system operating conditions must meet potential user expectations. The recognition vocabulary should cover a common lexicon and selected domains among, for instance, science, medicine, law etc. Our system covers an additional news domain (politics, economics, culture, education, sports and weather). Acoustically, the system is able to percept speech of every adequate user. In advance prepared speech, read text and spontaneous utterances can be recognized on similar level. User can record speech in conditions of home and office with widely available means. Naturally, significant noise and speaker overlaps should be avoided.

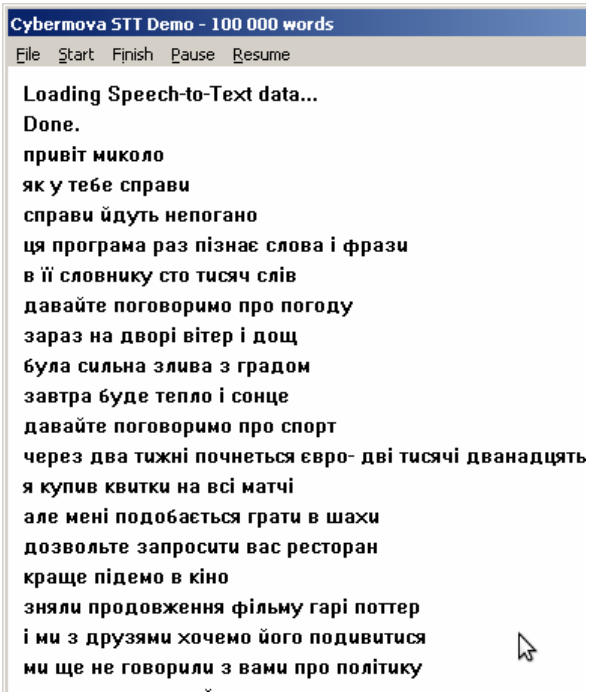

Figure 4. Dictation machine on PC demonstrates less than 5% WER on a general and news vocabulary speech segment of 94 words

In time of the system trial operation we varied its vocabulary size from 10 to 100 thousand words. Since the decoding process has been performed for all vocabularies in real time (less than 15% in *i*7 processor), more detail analyses was conducted for the largest vocabulary containing 100 000 words. The system has been tested as a dictation machine by ten experts. In described above operation conditions the word error rate did not exceed 10% on average. New word appending effectiveness was tested as well. Experts used to add proper names and rare terminology, which was introduced as *unknown* to the language model. The system ergonomics also benefited of abilities to put punctuation, paragraph breaks and undo the last operation using an isolated voice command.

4. Conclusion

This paper overviewed currently the most productive speech recognition scheme that applies the *analysis-by-synthesis* method. The system for Ukrainian speech-to-text conversion implementing the described algorithms shows its effectiveness through trial operation.

The development of the presented system is still on initial stage. In near future several improvements will be completed, which will increase accuracy and extend the system scope of usage. Among the improvements are: vocabulary enlargement, language model optimization by introducing word classes, context-dependent phoneme model application, speaker clusterization and adaptation, punctuation and character case prediction.

For both acoustic and language models, training set extension is important to represent entire speech and language variety. Therefore, adjustment of speech and text segment correspondence is crucial to facilitate thorough costly manual annotating procedure that could be even avoided at all under certain conditions. For text processing, more precise number and symbol to grapheme conversion is actual, particularly concerning to their function prediction and disambiguation.

For dictation machine, human-machine interaction becomes even more evolved. The system must suggest recognized utterance refinement based on multi-decision recognition response; moreover, accepted refinements must update the recognition response model. The function of recognition vocabulary extension is important for the user. Besides assigning a new word to the *unknown word* category, more sophisticated ways to update the vocabulary are to be implemented in order to keep the language model quality.

The more linguistic model is adequate to a subject domain, style and genre of speech the better is speech recognition accuracy. That is why to improve the system operation, text samples of the domain should be used adjust a language model that can be accomplished by means of basic text corpus interpolation.

5. References

[1]. http://www.forbes.com/sites/greatspeculations/2011/11/15/apple-trumps-google-on-voice-recognition-in-head-to-head-test

[2]. Винцюк Т.К., Анализ, распознавание и смысловая интерпретация речевых сигналов. Киев: Наукова думка, 1987, 264 с.

[3]. M. Gales and S. Young. "The Application of Hidden Markov Models in Speech Recognition." Foundations and Trends in Signal Processing, 2007, 1(3), pp. 195-304.

[4]. Taras Vintsiuk, Mykola Sazhok. Multi-Level Multi-Decision Models for ASR // Proceedings of the 10th Int. Conference on Speech and Computer – SpeCom'2005, Patras, 2005, pp. 69-76.

[5]. M. Gales. "Discriminative models for speech recognition" // ITA Work-shop, University San Diego, USA, February 2007.

[6]. G. Zweig. Speech Recognition with Dynamic Bayesian Networks. PhD thesis, University of California, Berkeley, 1998.

[7]. Робейко В.В., Сажок М.М. Багатозначна багато-рівнева модель перетворення орфографічного тексту на фонемний // Штучний інтелект. – № 4'2011. – Донецьк, 2011. – С. 117-125.

[8]. Робейко В.В., Сажок М.М. Використання текстового корпусу для прогнозування наголосів у словах української мови. // Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту: Матеріали міжнародної наукової конференції. – Херсон, 2012. – С. 171-172.

[9]. Young S.J. et al., The HTK Book Version 3.4, Cambridge University, 2006.

[10]. A. Lee, T. Kawahara. "Recent Development of Open-Source Speech Recognition Engine Julius" Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2009, pp. 131-137.

[11]. Bo-June (Paul) Hsu and James Glass. Iterative Language Model Estimation: Efficient Data Structure & Algorithms. In Proc. Interspeech, 2008

[12]. Н.Б. Васильєва, В.В. Пилипенко, О.М. Радуцький, В.В. Робейко, М.М. Сажок. Створення акустичного корпусу українського ефірного мовлення // Обробка сигналів і зображень та розпізнавання образів: X Міжнар. конференція УкрОбраз'2010. – Київ, 2010. – С. 55-58.

[13]. Широков В.А., Манако В.В. Організація ресурсів національної словникової бази // Мовознавство. – №5. – 2001 р. – С. 3–13.

[14]. www.cybermova.com/products/stt-demo.htm