NEWS
OF THENATIONAL ACADEMY OF SCIENCES OF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X **https://doi.org/10.32014/2020.2518-1726.64**
Volume 4, Number 332 (2020), 42 – 51

UDK 004.89
MRNTI 28.23.37

O. Mamyrbayev[1], D. Oralbekova[2]

[1]Institute of Information and Computational Technologies, Kazakhstan, Almaty;
[2]Satbayev University, Kazakhstan, Almaty.
E-mail: dinaoral@mail.ru

MODERN TRENDS IN THE DEVELOPMENT OF SPEECH RECOGNITION SYSTEMS

Abstract. This article presents the main ideas, advantages and disadvantages of models based on hidden Markov models (HMMs) - a Gaussian mixture models (GMM), end-to-end models and indicates that the end-to-end model is a developing area in the field of speech recognition. A review of studies that conducted in this subject area shows that end-to-end speech recognition systems can achieve results comparable to the results of standard systems using hidden Markov models, but using a simpler configuration and faster operation of the recognition system both in training and in decoding. An analytical review of the varieties of end-to-end systems for automatic speech recognition is considered, namely, models based on the connection time classification (CTC), attention-based mechanism and conditional random fields (CRF), and theoretical comparisons are made. Ultimately, their respective advantages and disadvantages and the possible future development of these systems are indicated.

Key words: automatic speech recognition, hidden Markov models, end-to-end, neural networks, CTC.

1. Introduction

Automatic speech recognition (ASR) is now widely used in our daily life. ASR can help people with disabilities interact with society. ASR is used in such areas as the automated user interface, mobile device management, information services and access control interfaces [1].

The purpose of ASR is identify the sequence of acoustic input $X = \{x_1, \cdots, x_T\}$ of length T as a sequence of words $W= \{w_1, \cdots, w_N\}$ of length N. The task of ASR is to find the most probable sequence of words W from given X. This can be represented as follows [2]:

$$W = \underset{W \in \gamma^*}{\operatorname{argmax}} * p(W \mid X). \quad (1)$$

Therefore, the main work of ASR is to create a model that can accurately calculate the posterior distribution p(W|X).

In the task of recognizing continuous and long speech, a model based on the Hidden Markov Model (HMM) was one of the best-known method. Even today, the best speech performance still comes from the HMM-based model combined with deep learning methods (hybrid models). At the same time, deep learning methods also simulated the emergence of an alternative, which is an end-to-end (E2E) model. This model, compared with HMM, uses one model to match directly sound to words. It replaces the design process with a learning process and does not require special knowledge in this area. Therefore, it is easier to create and train the E2E model. According to these advantages, the E2E model is quickly attracted much attention as a powerful method in the field of continuous speech recognition.

This article provides a detailed overview of the E2E model, as well as a brief comparison between the HMM-based model and the E2E model, an analysis of the various paradigms of E2E models and a comparison of their advantages and disadvantages. First, consider the main methods of speech recognition.

2. The main methods of speech recognition

2.1 Speech Processing Methods

Currently, there are several basic approaches for ASR.

The standard process for automatic speech recognition consists of the following steps:

– Feature extraction from the input signal.
– Acoustic modeling.
– Language modeling.
– Decoding sequence.

The most important parts of a speech recognition system are feature extraction methods and recognition methods. Feature extraction is a process that extracts a small amount of data from a signal [4]. At the beginning, the original signal is converted into feature vectors, based on which classification will then be performed. This step includes the following steps:

– conversion of the signal into digital form;
– the use of various filters to suppress noise;
– highlighting the boundaries of speech;
– extraction of signal features [5].

The most popular extraction features methods are the Mel Frequency Cepstral Coefficients (MFCC) and the linear prediction cepstral coefficients (PLP). MFCC is an audio function extraction method that extracts the speaker's specific parameters from speech [6]. MFCCs are extracted from speech signals through cepstral analysis. The input signal is first formed and processed in the form of a window, then the Fourier transform is taken and the value of the resulting spectrum is deformed according to the Mel scale [7].

By using the obtained feature vectors, it is necessary to determine which sound or sequence of words was in the original signal. Widespread methods of automatic speech recognition (ASR) are hidden Markov models (HMM) and neural networks (NN) [5].

2.2 Standard Speech Recognition System. HMM-based model

For a long time, the HMM-based model was the main model for continuous speech recognition with a large dictionary with better recognition results. In general, an HMM-based model can be divided into three parts; each of them is independent of each other and plays a different role: acoustic, pronunciation and language model. The acoustic signal of speech is modeled by a small set of acoustic units, which can be considered as elementary sounds of the language. The traditionally chosen unit is a phoneme, so the word is formed by combining them [8]. The pronunciation model, which is usually created by professional human linguists, is used to achieve a correspondence between phonemes (or sub-phonemes) and graphemes. The language model maps a sequence of characters into free final transcription [9].

The HMM mechanism was used in all of these three parts. However, an HMM-based model emphasized the use of HMM in an acoustic model. In this HMM, sound was observation, and feature was a latent state. For an HMM that had a set of states $\{1, \cdots, J\}$, the HMM-based model used the Bayesian theorem and introduced the sequence of states HMM $S = \{s_t \in \{1, \cdots, J\} \mid t = 1, \cdots, T\}$ and expanded $p(L \mid X)$.

$$\begin{aligned} \underset{L \in \gamma^*}{argmax}\, p(L|X) &= \underset{L \in \gamma^*}{argmax}\, \frac{p(L,X)}{P(X)} \\ &= \underset{L \in \gamma^*}{argmax}\, p(L,X) \\ &= \underset{L \in \gamma^*}{argmax} \sum_S p(P,L,X) \\ &= \underset{L \in \gamma^*}{argmax} \sum_S p(X|S,L)\, p(S,L) \\ &= \underset{L \in \gamma^*}{argmax}\, \textstyle\sum_S p(X|S,L)\, p(S|L)\, p(L) \end{aligned} \quad (2)$$

According to the conditionally independent hypothesis, we can approximate p (X | S, L) ≈p (X | S), therefore

$$\underset{L \in \gamma^*}{argmax}\, p(L|X) \approx \underset{L \in \gamma^*}{argmax} \sum_S p(X|S)\, p(S|L)\, p(L) \quad (3)$$

p (X | S), p (S | L), and p (L) in equation (3) correspond to the acoustic model, pronunciation model, and language model, respectively.

– The acoustic model P (X | S) indicates the probability of observing X from a hidden sequence S. According to the rule of the chain of probabilities and the hypothesis of independence of observations in HMM (observations at any time depend only on the latent state at that time), P (X | S) can be laid out in the following form:

$$p(X|S) = \prod_{t=1}^{T} p(x_t|x_1, \ldots, x_{t-1}, S) \approx \prod_{t=1}^{T} p(x_t|s_t) \propto \prod_{t=1}^{T} \frac{p(s_t|x_t)}{p(s_t)} \quad (4)$$

In the acoustic model p (x_t | s_t) is the probability of observation, which is usually represented by the Gaussian Mixture Model (GMM). The distribution of the posterior probability of the latent state p (s_t | x_t) can be calculated using the method of deep neural networks (DNN). These two different calculations of P (X | S) lead to two different models, namely HMM-GMM and HMM-DNN. Over time, the HMM-GMM model has been a common framework for speech recognition. With the development of deep learning technology, DNN is being introduced into speech recognition for acoustic modeling. The role of DNN is to calculate the posterior probability of the state of the HMM, which can be converted into probability, replacing the usual probability of observing GMM [10]. Thus, the HMM-GMM model turns into an HMM-DNN, which achieves better results than the HMM-GMM, and becomes a modern ASR model.

In an HMM-based model, different modules use different technologies and play different roles. HMM is mainly used for dynamic time warping at the frame level. GMM and DNN are used to calculate the probability of emission of latent HMM states. The building process and the mode of operation of the model based on HMM determines whether they encounter the following difficulties in practical use [11]:

– The training process is complex and difficult for global optimization. An HMM-based model often uses different training methods and data sets to train different modules. Each module is independently optimized using its own target optimization functions, which usually differ from the true criteria for evaluating the performance of continuous speech recognition. Thus, the optimality of each module does not necessarily mean global optimality.

– Conditionally independent assumptions. To simplify model building and training, an HMM-based model uses assumptions about conditional independence within HMM and between different modules.

2.3 End-to-end ASR models

End-to-end (E2E) automatic speech recognition is a new paradigm in the field of speech recognition based on a neural network, which offers many advantages. Traditional "hybrid" ASR systems, which consist of an acoustic model, a language model, and a pronunciation model, require separate training for these components, each of which can be complex. For example, training an acoustic model is a multi-stage process of training a model and aligning the time between a sequence of acoustic characteristics of speech and a sequence of labels at the output. The E2E ASR, by contrast, is a single integrated approach with a much simpler learning pipeline with models that work with low audio frame rates. This reduces training time, decoding time and allows joint optimization with subsequent processing, such as understanding of a natural language.

However, modern E2E ASR systems also have some limitations:

Firstly, E2E ASR systems require more training data than ASR hybrid systems to achieve a similar word error rate (WER). This is because E2E ASR systems tend to exceed training data when they are limited.

Secondly, Connectionist Temporal Classification (CTC), a popular version of the E2E ASR, is not amenable to 'student-teacher' training, which is useful for deploying high-precision ASR systems with time-out limits [12].

The end-to-end model can be divided into three different categories depending on their smooth alignment implementations: CTC, attention-Based Models, the model, based on Conditional Random Fields (CRF).

2.3.1 End-to-end model based on Connectionist Temporal Classification

Although the HMM-DNN hybrid model still has the most up-to-date results, the role of DNN is limited. It is mainly used to model the probability of an a posteriori state of the latent state of HMM, presenting only local information. The temporary domain function is still modeled by HMM. By trying to simulate objects in the time domain using RNN or convolutional neural networks (CNN) instead of HMM, he encounters the problem of data alignment. The loss functions of both RNN and CNN (Convolutional Neural Networks) are determined at each point in the sequence, therefore, to provide training opportunities, need to know the alignment relationship between the output RNN sequence and the target sequence [13].

The CTC process can be thought of as including two subprocesses: calculating the probability of a path and aggregating a path. In these two subprocesses, the most important is the introduction of a new blank label ("-", which means no output) and an intermediate path to the concept.

By solving these two problems, CTC can use a single network structure to map the input sequence directly to the label sequence and implement end-to-end speech recognition.

For a given input sequence $X = \{ x_1, \cdots, x_T \}$ of length T, the encoder encodes it into a sequence of signs $F = \{ f_1, \cdots, f_T \}$ of length T for any t, f_t - this is a vector whose dimension is greater than the number of elements in the dictionary γ, i.e., $f_t \in R^{|\gamma|+1}$.

CTC acts on the sequence of signs $F = \{ f_1, \cdots, f_T \}$. Through the softmax operation, CTC transforms it into a probability distribution sequence $Y = \{ y_1, \cdots, y_T \}$, $yt = \{ y^1_t, \cdots, y_t^{|\gamma+1|} \}$, where y^1_t indicates the probability that the output signal at time step t is the label i, $y_t^{|\gamma+1|}$ indicates the probability of outputting an empty label at time step t.

Let $\gamma' = \gamma \cup \{b\}$, γ'^T denote the set of all sequences of length T defined in the dictionary γ′. In combination with the definition of y^k_t, we can conclude that for a given input sequence X, the conditional probability distribution of any sequence π in the set γ'^T is calculated as equation (5):

$$p(\pi|X) = \prod_{t=1}^{T} y_t^{\pi t}, \forall \pi \in \gamma'^T \quad (5)$$

where π_t represents the label at position t of the sequence π. An element in γ'^T is a path and is represented as π.

After the calculation process described above, the input sequence $\{x_1, \cdots, x_T\}$ is mapped onto a path π of the same length, and the conditional probability π can also be calculated in accordance with equation (5). In this mapping process, each input x_t frame is mapped to a specific label π_t. This might be thought that mapping an input sequence to a path is actually a tightly coordinated process.

From the process of calculating equation (5), we can see that there is a very important assumption, which is an assumption of independence: the elements in the output sequence are independent of each other. Any time step whose label is selected as the output does not affect the distribution of marks at other time steps. On the contrary, in the coding process, the value of y^k_t is influenced by the speech context information in both historical and future directions. That is, CTC uses conditional independence conditions in language models, but not in acoustic models. Therefore, the encoder obtained by training CTC is essentially and completely an acoustic model that is not capable of modeling the language.

Let $\gamma \leq T$ denote the set of all label sequences defined in the dictionary γ whose length is less than or equal to T, and path aggregation is defined as a function of the map O: $L'T \mapsto L \leq T$. It maps paths in γ'^T (that is, a path) to a real label sequence in $\gamma^{\leq T}$. Aggregation of O paths mainly consists of two operations:

1. The union of the same adjacent labels. If consecutive identical marks appear in the path, combine them and leave only one of them. For example, for two different paths "d-oo-t-" and "d-o-tt-", they are aggregated in accordance with the above principles to obtain the same result: "d-o-t-".

2. Removing the empty "-" mark in the path. Since the "-" label indicates no output, it should be deleted when the final label sequence is generated. The above sequence "d-o-t-" after aggregation in accordance with this principle becomes the final sequence "dot".

In addition to obtaining a sequence of labels corresponding to these paths, aggregation is also aimed at calculating the probability of a sequence of labels. We use O^{-1} (L) to represent the set of all paths in γ'^T corresponding to the sequence of labels L, then, obviously, given the input sequence X, the probability p (L | X) for L can be calculated as in equation (6):

$$p(L|X) = \sum_{\pi \in O^{-1}(L)} p(\pi|X) \quad (6)$$

Obviously, the calculation of the probability L is differentiable. Therefore, after obtaining the probability of the label for training the model, it can be used the back propagation method of the error.

However, there is still difficulty in calculating equation (6). Although p (π | X) is easy to calculate, it is difficult to determine which and how many paths from γ^{T} are included in O^{-1} (L). Consequently, this equation is not actually used to calculate p (L | X). Its operational method of calculation is the forward and reverse algorithm.

The advent of CTC technology greatly simplifies the design and training of continuous speech recognition models. No longer required experience to create various dictionaries; this eliminates the need for data alignment, allowing us to use any number of layers, any network structure to build an end-to-end model that maps sound directly to text [14].

One of the great benefits of CTC is that it eliminates the need to align data segmentation, so deep learning techniques such as CNN and RNN can play an increasingly important role. Network models with different structure and depth were introduced in the end-to-end ASR and achieved better results.

2.3.2 Attention-Based Model

An alternative approach to the end-to-end mapping between speech and tag sequences is to use an encoder-decoder architecture based on the attention mechanism [15]. This architecture has two separate subnets. One of them is the encoder subnet, which converts the sequence of acoustic features into a sequential representation of the length T. Based on this encoded information, the decoder subnet predicts a sequence of labels whose length L is usually less than the input length. The decoder uses only the relevant portion of coded sequential representations to predict the label at each time step using the attention mechanism.

The encoder is implemented as a multilayer bidirectional recurrent neural network (RNN), such as Long short-term memory (LSTM), and the decoder usually consists of a 1st level unidirectional RNN, followed by the output layer softmax. The structure of the attention-based model is shown in figure 1 [16].

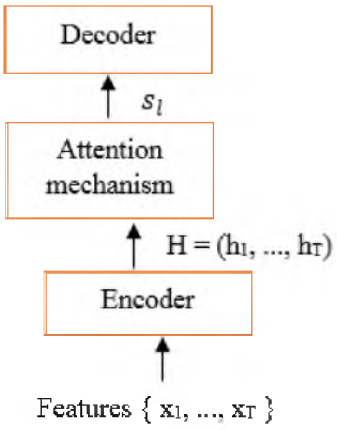

Figure 1 - Attention Mechanism Model

The attention-based model is formulated as follows. The encoder converts X into intermediate representation vectors H = (h_1, ..., h_T). At the next stage of decryption, activation of the latent state (memory) of the RNN-based decoder at the l-th time step is calculated as:

$$s_l = Recurrency(s_{l-1}, g_l, y_{l-1}) \quad (7)$$

where g_l and y_{l-1} denote a "glimpse" at the l-th time step and the predicted mark at the previous step. The glimpse g_l is a weighted sum of the encoder output sequence as

$$g_l = \sum_l \alpha_{l,t} h_t \quad (8)$$

where $\alpha_{l,t}$ – weight of attention h_t and calculated as

$$e_{l,t} = Score(s_{l-1}, h_t, \alpha_{l-1}) \quad (9)$$

$$\alpha_{l,t} = \frac{\exp(e_{l,t})}{\sum_{t'=1}^{T} \exp(e_{l,t})} \quad (10)$$

The encoder-decoder method that use the attention mechanism does not require preliminary data segmentation. With attention, it can implicitly learn the soft alignment of input and output sequences, which solves a big problem for speech recognition [17].

The encoder plays the role of an acoustic model, which is the same as in CTC models, RNN converters, and even hybrid HMM-DNN models. Thus, he faces the same problems as they, and their solutions are the same. However, when the encoder is combined with attention, new problems arise [18].

A serious problem caused by the combination of encoder and attention is delay. Since attention is paid to the entire sequence of coding results, it is necessary to wait until the encoding process is fully completed before it can start working, so the time spent on the encoding process will increase the model delay. In addition, an encoder that does not reduce the length of the sequence will have a sequence of encoding results that is much longer than the target label sequence (for the input speech sequence is much longer than for transcription) [19]. This leads to two problems: on the one hand, a longer sequence of encoding results means more attention, thereby increasing the delay; on the other hand, since speech is much more than transcription, the sequence generated by the encoding process without sub-sampling that will bring a lot of redundant information to the attention mechanism.

Similar to the development trend in models based on CTC and RNN converters, to improve the encoding capabilities, the encoder in attention-based models is also becoming more and more complex. The most obvious moment is reflected in its depth. The early encoder was mainly in three layers and gradually developed to six layers. As the network structure becomes more complex and its depth deepens, the model effect is constantly improving. In [20], a 15-layer network of encoders was built by using network on the network, packet normalization, residual network, convolutional LSTM, and ultimately achieved a WER of 10.53% without using a dictionary or language model.

2.3.3 Conditional Random Field Model

Conditional Random Fields (CRF) model allows to combine local information to predict the global probabilistic model for sequences. This model was first proposed in [21] for speech recognition.

In this method, X is a random variable for the data sequences that is labeled, and Y is a random variable for the corresponding label sequences. All Y_i components from Y are located in the alphabet of the final label Y. Random variables X and Y are distributed together, but in the discriminatory structure they must build a conditional model p (Y | X) from pair observations and sequences of labels. Let G = (V, E) be a graph, and $Y = (Y_v)_{v \in V}$, so that Y is indexed by the vertices of G. Then (X, Y) is a conditional random field in the case where the condition on X is random variables Y_v obey the Markov property with respect to the graph: p (Y_v | X, Y_w, w ≠v) = p (Y_v | X, Y_w, w ~ v), where w ~ v means that w and v are neighbors in G. The structure of the model on CRF-based is presented in figure 2.

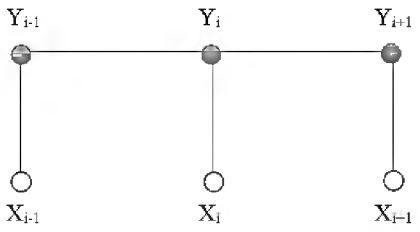

Figure 2 - Graphical representation of a CRF model

The most common application is the linear chain CRF model. This model is most often used to solve the problems of marking and segmentation of sequences [22].

A similar method for CRF is the MEMM algorithm (maximum-entropy Markov model), which is also a discriminative probabilistic model. The main difference between CRF and MEMM is the absence of a label bias problem (label bias is a situation where states with fewer transitions take precedence, since a single probability distribution and normalization are built) [23].

According to [24], [25] studies, after using CRF, better results were obtained than MEMM or HMM without using a language model.

3. Conclusion

The considered methods for constructing end-to-end models are superior the HMM-GMM model, but its performance is still worse or comparable to the HMM-DNN model, which also uses methods of deep learning. In order to take an advantage of the end-to-end model, there should be at least improved in the following aspects:

– CTC-based models are monotonous and support stream decoding, so they are suitable for low-latency online scenarios. However, their recognition efficiency is limited. The main disadvantage of the CRF-based model is the computational complexity of the training sample analysis, which makes it difficult to constantly update the model when new training data arrives. Models based on the attention mechanism can effectively improve recognition characteristics, but they are not monotonous and have a long delay. Although methods exist such as 'a window' to reduce attention delay, they can reduce recognition performance to some extent. Therefore, reducing latency while ensuring performance is an important research problem for the end-to-end model.

– The HMM-based model uses additional language models to provide a wealth of language training, while all linguistic training of the end-to-end model is obtained only from transcriptions of training data, the scope of which is very limited. This leads to great difficulties when working with scenes with great linguistic diversity. Therefore, the E2E model should improve the study of linguistic training while maintaining the integral structure. This article was prepared based on the project: IRN AP05131207 Development of technology for multilingual automatic speech recognition using deep neural networks.

О.Ж. Мамырбаев[1], Д.О. Оралбекова[2]

[1]Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан;
[2]Satbayev University, Алматы, Қазақстан

СӨЙЛЕУДІ ТАНУ ЖҮЙЕСІНІҢ ДАМУЫНДАҒЫ ҚАЗІРГІ ТЕНДЕНЦИЯЛАР

Аннотация. Бұл мақалада жасырын Марков модельдеріне (HMMs) негізделген модельдердің негізгі идеялары, артықшылықтары мен кемшіліктері - Gaussian үлестірімдері (GMM) және интегралдық жүйелер (end-to-end) гибриді ұсынылған, сонымен қатар интегралды модель сөйлеуді тану саласында дамып келе жатқан жаңа саланың бірі болып табылады.

Кіріккен сөйлеуді тану проблемасында жасырын Марков моделіне (HMM) негізделген модель әрдайым басты технология болып келді және кең қолданылды. Тінті қазіргі уақытта сөйлеуді тану бойынша ең жақсы көрсеткіш HMM-ге негізделген (терең оқыту әдістерімен біріктірілген). Көптеген өнеркәсіптік орналастырулар HMM-ге негізделген.

Сонымен бірге терең оқыту әдістері интегралды модель болып табылатын баламаның пайда болуына түрткі болды. HMM негізіндегі модельмен салыстырганда дыбысты таңбаларга немесе сөздерге тікелей сәйкестендіру үшін біріктірілген модель бір модельді қолданады. Ол өңдеу процесін оқыту үрдісімен алмастырады және осы салада арнайы оқытуды қажет етпейді, сондықтан интегралды модельді құру және оқыту оңайырақ. Осы артықшылықтардың арқасында интегралды модель тез арада сөйлеуді тану саласындағы танымал зерттеу аймагына айналып отыр.

Көптеген интегралды сөйлеуді тану модельдеріне келесі бөліктер кіреді: сөйлеу енгізу тізбегін мүмкіндіктер тізбегімен салыстыратын кодер; объектілер тізбегі мен тіл арасындағы теңестіруді жүзеге асыратын түзету; түпнұсқалық сәйкестендіру нәтижесін ашатын декодер. Тағы айта кететін жайт, бұл бөлу әрдайым бола бермейді, өйткені интегралдық құрылымның өзі толық құрылым болып табылады және инженерлік модульдік жүйеге ұқсастыгы бойынша жұмысты қай бөлігі орындайтынын анықтау өте қиын.

Бірнеше модульден тұратын HMM моделінен айырмашылыгы, интегралды модель акустикалық сигналдардың тікелей көрсетілуін іске асыратын терең желілер модулін алмастырады. Сонымен қатар, нәтиже шыгарган кезде кейінгі өңдеуді қажет етпейді.

HMM негізделген модельмен салыстырганда, жогарыдагы айырмашылықтар интегралды модельге келесі сипаттамаларды береді:

– бірлескен жаттыгулар үшін бірнеше модульдер бір желіге біріктірілген. Бірнеше модульді біріктірудің артықшылыгы - әр түрлі аралық жағдайының (қалпының) айырмашылыгы арасындагы көрсетілуін жүзеге асыру үшін көп модуль жасаудың қажеттігі жоқ. Бірлескен оқыту интегралды модельге жаһандық

оңтайландыру мақсаты ретінде қорытынды бағалау критерийлері үшін өте маңызды функцияны қолдануға мүмкіндік береді, осылайша тиімді нәтижелерге қол жеткізуге мүмкіндік береді.

– акустикалық сигнатурасы мәтіндік нәтижесінің тізбегіне тікелей сәйкестендіріледі және нақты транскрипцияға қол жеткізу немесе тану сипаттамаларын жақсарту үшін одан әрі өңдеуді қажет етпейді, ал HMM модельдерінде айтылым үшін ішкі оқыту бар.

Интегралды модельдің осы артықшылықтары сөйлеуді тану модельдерінің құрылысы мен жаттығуын айтарлықтай жеңілдетеді.

Интегралды модельді олардың жұмсақ теңестірілуіне байланысты үш санатқа бөлуге болады:

- CTC негізінде: CTC алдымен барлық ықтимал теңестірулерді тізімдейді, содан кейін осы қатты туралауларды біріктіріп жұмсақ теңестіруге қол жеткізеді. CTC шығыс белгілері қатаң туралауды тізімдеген кезде бір-бірінен тәуелсіз болады деп болжайды.

- назар аудару механизміне негізделген: бұл әдіс кіріс деректері мен шығыс белгілері арасындағы тегістеу ақпаратын тікелей есептеу үшін назар аудару механизмін қолданады.

- CRF шартты кездейсоқ өрістеріне негізделген модель ғаламдық ықтималды модельді тізбектей болжау үшін жергілікті ақпаратты біріктіруге мүмкіндік береді.

Осылайша, сөйлеуді автоматты түрде тануға арналған интегралды жүйелердің түрлеріне аналитикалық шолу және теориялық салыстырулар жасалды. Соңында, олардың тиісті артықшылықтары мен кемшіліктері және осы жүйелердің болашақта дамуы мүмкін перспективасы көрсетілген. HMM моделі мен интегралды модель туралы қысқаша салыстырамыз. Сайып келгенде, олардың тиісті артықшылықтары мен кемшіліктері және осы жүйелердің болашақта дамуы мүмкін екендігі көрсетілген.

Түйін сөздер: сөйлеуді автоматты түрде тану, жасырын Марков модельдері, end-to-end; нейрондық желілер, CTC.

О.Ж. Мамырбаев[1], Д.О. Оралбекова[2]

[1]Институт информационных и вычислительных технологий, Казахстан, Алматы;
[2]Satbayev University, Казахстан, Алматы

СОВРЕМЕННЫЕ ТЕНДЕНЦИИ РАЗВИТИЯ СИСТЕМ РАСПОЗНАВАНИЯ РЕЧИ

Аннотация. В данной статье представлены основные идеи, преимущества и недостатки моделей, на основе скрытых марковских моделей (HMM) - смеси гауссовских распределений (GMM) и интегральных систем (end-to-end), а также указано, что интегральная модель является развивающим направлением в области распознавания речи.

В задаче распознавания слитной речи широко использовалась модель на основе скрытой модели Маркова (HMM) и всегда была одной из основных технологий в этой области. Даже сегодня лучшая производительность распознавания речи по-прежнему исходит от модели на основе HMM (в сочетании с методами глубокого обучения). Большинство промышленно развернутых систем основаны на HMM.

В то же время, методы глубокого обучения также стимулировали появление альтернативы, которая является интегральной моделью. По сравнению с моделью, основанной на HMM, в интегральной модели используется одна модель для непосредственного сопоставления звука с символами или словами. Он заменяет процесс проектирования процессом обучения и не требует специальных знаний в этой области, поэтому интегральную модель проще создавать и обучать. Благодаря этим преимуществам интегральная модель быстро становится популярным направлением исследований в области распознавания речи.

Большинство интегральных моделей распознавания речи включают в себя следующие части: кодер, который отображает последовательность ввода речи в последовательность признаков; выравниватель, который реализует выравнивание между последовательностью объектов и языком; декодер, который декодирует окончательный результат идентификации. Необходимо отметить, что это разделение не всегда существует, потому что интегральная структура сама по себе является законченной структурой, и обычно очень трудно определить, какая часть выполняет какую подзадачу по аналогии с инженерной модульной системой.

В отличие от модели на основе HMM, которая состоит из нескольких модулей, интегральная модель заменяет несколько модулей глубокой сетью, реализуя прямое отображение акустических сигналов в последовательности меток без тщательно продуманных промежуточных состояний. Кроме того, нет необходимости выполнять последующую обработку на выходе.

По сравнению с моделью на основе HMM вышеуказанные различия дают интегральной модели следующие преимущества:

– несколько модулей объединены в одну сеть для совместного обучения. Преимущество объединения нескольких модулей состоит в том, что нет необходимости разрабатывать много модулей для реализации

отображения между различными промежуточными состояниями. Совместное обучение позволяет интегральной модели использовать функцию, которая очень важна для окончательных критериев оценки, в качестве цели глобальной оптимизации, тем самым добиваясь глобально оптимальных результатов.

– он напрямую отображает входную последовательность акустической сигнатуры в последовательность текстового результата и не требует дальнейшей обработки для достижения истинной транскрипции или для улучшения характеристик распознавания, тогда как в моделях на основе HMM обычно есть внутреннее представление для произношения.

Эти преимущества интегральной модели позволяют значительно упростить построение и обучение моделей распознавания речи.

Интегральная модель может быть разделена на три различные категории в зависимости от их реализаций гладкого выравнивания:

– на основе CTC: CTC сначала перечисляет все возможные грубые выравнивания, затем он достигает гладкого выравнивания путем объединения этих грубых выравниваний. CTC предполагает, что выходные метки не зависят друг от друга при перечислении таких выравниваний.

– основанная на механизме внимания: этот метод использует механизм внимания, чтобы непосредственно вычислить информацию гладкого выравнивания между входными данными и выходной меткой.

– модель, на основе условных случайных полей (Conditional Random Fields, CRF), позволяет комбинировать локальную информацию для прогнозирования глобальной вероятностной модели по последовательностям.

Таким образом, в статье приведен аналитический обзор разновидностей интегральных систем автоматического распознавания речи, и делаются теоретические сравнения. Кратко сравниваются модели на основе HMM и интегральная модель, тщательно анализируются различные парадигмы интегральных технологий и сравниваются их преимущества и недостатки. В конечном итоге указываются возможное будущее развитие этих систем.

Ключевые слова: автоматическое распознавание речи, скрытые марковские модели, end-to-end; нейронные сети, CTC.

Information about authors:

Mamyrbayev Orken, ass. Professor, PhD-doctor of information systems, Institute of Information and Computational Technologies, morkenj@mail.ru, https://orcid.org/0000-0001-8318-3794;

Oralbekova Dina Orymbayevna, PhD-student, Satbayev University, dinaoral@mail.ru, https://orcid.org/0000-0003-4975-6493

REFERENCES

[1] Kazachkin A. E. **(2019)** Speech recognition methods, modern speech technology [Metody raspoznavanija rechi, sovremennye rechevye tehnologii] // Young scientist. №39. pages 6-8. URL https://moluch.ru/archive/277/62675/ (accessed: 28.01.2020)]. (In Russian).

[2] Ronzhin A.L., Karpov A.A., Li I.V. **(2006)** Voice and multimodal interfaces [Rechevoj i mnogomodal'nyj interfejsy] // M.: Science. Page 173] (In Russian).

[3] Gusev M.N. **(2013)** Speech Recognition System: basic models and algorithms [Sistema raspoznavanija rechi: osnovnye modeli i algoritmy] / M.N. Gusev, V.M. Degtjarev. SPb.: Sign, page 128 c.] (In Russian).

[4] Ibrahim M. El-Henawy, Walid I. Khedr, Osama M. ELkomy, Al-Zahraa M.I. AbdallaЮ **(2014)** Recognition of phonetic Arabic figures via wavelet based Mel Frequency Cepstrum using HMMs, HBRC Journal, Volume 10, Issue 1, 2014, Pages 49-54, ISSN 1687-4048. DOI:10.1016/j.hbrcj.2013.09.003 (in Eng).

[5] Vorob'eva S. A. **(2016)** Speech Recognition Methods [Metody raspoznavanija rechi] // Young scientist. №26. pages 136-141. URL https://moluch.ru/archive/130/36213/ (accessed: 28.01.2020). (In Russian).

[6] Sirko Molau, Michael Pitz, Ralf Schluter and Hermann Ney. **(2001)** "Computing Mel frequency Cepstral Coefficients on the power spectrum." IEEE Transactions on Audio, Speech and Language Processing. DOI: 10.1109/ICASSP.2001.940770 (in Eng).

[7] Bezoui Mouaz, Beni Hssane Abderrahim, Elmoutaouakkil Abdelmajid. **(2019)** Speech Recognition of Moroccan Dialect Using Hidden Markov Models, Procedia Computer Science, Volume 151, Pages 985-991, ISSN 1877-0509. DOI:10.11591/ijai.v8.i1.pp7-13 (in Eng).

[8] Rabiner L-R., Juang B-H. **(1993)** Fundamentals of Speech Recognition, Prentice-Hall. Englewood Cliffs, N.J.: PTR Prentice Hall, 1993. (in Eng).

[9] Rao, K.; Sak, H.; Prabhavalkar, R. **(2017)** Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan; pp. 193–199.]. (in Eng).

[10] Lu, L.; Zhang, X.; Cho, K.; Renals, S. **(2015)** A study of the recurrent neural network encoder-decoder for large vocabulary speech recognition. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany; pp. 3249–3253. (in Eng).

[11] Rahhal Errattahi, Asmaa El Hannani, Hassan Ouahmane. **(2018)** Automatic Speech Recognition Errors Detection and Correction: A Review, Procedia Computer Science, Volume 128, Pages 32-37, ISSN 1877-0509, DOI: 10.1016/j.procs.2018.03.005 (in Eng).

[12] Mamyrbayev O., Alimhan K., Zhumazhanov B., Turdalykyzy T., Gusmanova F. **(2020)** End-to-End Speech Recognition in Agglutinative Languages. In: Nguyen N., Jearanaitanakij K., Selamat A., Trawiński B., Chittayasothorn S. (eds) Intelligent Information and Database Systems. ACIIDS 2020. Lecture Notes in Computer Science, vol 12034. Springer, Cham. DOI: 10.1007/978-3-030-42058-1_33 (in Eng).

[13] O Mamyrbayev, A Toleu, G Tolegen, N Mekebayev. **(2020)** Neural architectures for gender detection and speaker identification. Cogent Engineering 7 (1), 1727168. (in Eng).

[14] O Mamyrbayev, N Mekebayev, M Turdalyuly, N Oshanova, TI Medeni. **(2019)** Voice Identification Using Classification Algorithms. Intelligent System and Computing. DOI: 10.5772/intechopen.88239 (in Eng).

[15] Ueno, Sei & Inaguma, Hirofumi & Mimura, Masato & Kawahara, Tatsuya. **(2018)** Acoustic-to-Word Attention-Based Model Complemented with Character-Level CTC-Based Model. 5804-5808. 10.1109/ICASSP.2018.8462576.]. DOI: 10.1109/ICASSP.2018.8462576 (in Eng).

[16] Prabhavalkar, R.; Rao, K.; Sainath, T.N.; Li, B.; Johnson, L.; Jaitly, N. **(2017)** A comparison of sequence-to-sequence models for speech recognition. In Proceedings of the Interspeech, Stockholm, Sweden; pp. 939–943. (in Eng).

[17] M.N. Kalimoldayev, O.Z. Mamyrbayev, A.S. Kydyrbekova, N.O. Mekebayev. **(2019)** Voice verification and identification using i-vector representation. International Journal of Mathematics and Physics 10 (1), 66-74. DOI: 10.26577/ijmph-2019-i1-9 (in Eng).

[18] Wang, Dong & Wang, Xiaodong & Lv, Shaohe. **(2019)** An Overview of End-to-End Automatic Speech Recognition. Symmetry. 11. 1018. 10.3390/sym11081018. DOI:10.3390/sym11081018 (in Eng).

[19] O.Zh.Mamyrbayev, M.Turdalyuly, N.O.Mekebaev, A.S.Kydyrbekova. **(2019)** "Automatic Recognition of the Speech Using Digital Neural Networks", ACIIDS, Indonesia, Proceedings , Part II. (in Eng).

[20] Bahdanau, D.; Chorowski, J.; Serdyuk, D.; Brakel, P.; Bengio, Y. **(2016)** End-to-end attention-based large vocabulary speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China; pp. 4945–4949.] DOI: 10.1109/ICASSP.2016.7472618 (in Eng).

[21] J. Lafferty, A. McCallum, and F. Pereira. **(2001)** "Conditional random fields: Probabilistic models for segmenting and labeling sequence data" in Proceedings of the International Conference on Machine Learning (ICML'01), Williamstown, MA, USA, pp. 282–289. (in Eng).

[22] E. Fosler-Lussier, Y. He, P. Jyothi, and R. Prabhavalkar. **(2013)** "Conditional random fields in speech, audio, and language processing" Proceedings of the IEEE, vol. 101, no. 5, pp. 1054–1075. DOI: 10.1109/JPROC.2013.2248112 (in Eng).

[23] N. M. Markovnikov, I. S. Kipjatkova. **(2018)** Analytical review of end-to-end speech recognition systems [Analiticheskij obzor integral'nyh sistem raspoznavanija rechi], SPIIRAN, 58, pages 77–110 (In Russian).

[24] Hifny Y., Renals S. **(2009)** Speech recognition using augmented conditional random fields // IEEE Transactions on Audio, Speech, and Language Processing, vol. 17. no. 2, pp. 354–365. (in Eng).

[25] H. Tang et al.**(2017)** "End-to-End Neural Segmental Models for Speech Recognition," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1254-1264. DOI: 10.1109/JSTSP.2017.2752462 (in Eng).