JOURNAL OF SCIENCE

SAKARYA UNIVERSITY

Sakarya University Journal of Science

ISSN 1301-4048 | e-ISSN 2147-835X | Period Bimonthly | Founded: 1997 | Publisher Sakarya University | http://www.saujs.sakarya.edu.tr/en/

Title: Deep Learning Based Automatic Speech Recognition For Turkish

Authors: Burak TOMBALOĞLU, Hamit ERDEM

Recieved: 2020-04-02 00:24:21

Accepted: 2020-05-28 10:25:30

Article Type: Research Article
Volume: 24
Issue: 4
Month: August
Year: 2020
Pages: 725-739

How to cite
Burak TOMBALOĞLU, Hamit ERDEM; (2020), Deep Learning Based Automatic Speech Recognition For Turkish. Sakarya University Journal of Science, 24(4), 725-739,
DOI: https://doi.org/10.16984/saufenbilder.711888
Access link
http://www.saujs.sakarya.edu.tr/en/pub/issue/55932/711888

JOURNAL of SCIENCE
Sakarya University
Journal of Science
VOLUME: 24
ISSUE: 4
AUGUST 2020
e-issn: 2147-835X

Sakarya University Journal of Science 24(4), 725-739, 2020

SAKARYA
UNIVERSITY

Deep Learning Based Automatic Speech Recognition for Turkish

Burak TOMBALOĞLU*1, Hamit ERDEM2

Abstract

Using Deep Neural Networks (DNN) as an advanced Artificial Neural Networks (ANN) has become widespread with the development of computer technology. Although DNN has been applied for solving Automatic Speech Recognition (ASR) problem in some languages, DNN-based Turkish Speech Recognition has not been studied extensively. Turkish language is an agglutinative and a phoneme-based language. In this study, a Deep Belief Network (DBN) based Turkish phoneme and speech recognizer is developed. The proposed system recognizes words in the system vocabulary and phoneme components of out of vocabulary (OOV) words. Sub-word (morpheme) based language modelling is implemented into the system. Each phoneme of Turkish language is also modelled as a sub-word in the model. Sub-word (morpheme) based language model is widely used for agglutinative languages to prevent excessive vocabulary size. The performance of the suggested DBN based ASR system is compared with the conventional recognition method, GMM (Gaussian Mixture Method) based Hidden Markov Model (HMM). Regarding to performance metrics, the recognition rate of Turkish language is improved in compare with previous studies.

Keywords: Deep Neural Networks, Automatic Speech Recognition, Deep Belief Networks, Turkish

*Corresponding Author: buraktombaloglu@hotmail.com
[1] Başkent University, Electrical and Electronics Engineering, Ankara, Turkey.
ORCID: https://orcid.org/0000-0003-3994-0422
[2] Başkent University, Electrical and Electronics Engineering, Ankara, Turkey.
ORCID: https://orcid.org/0000-0003-1704-1581. E-mail: herdem@baskent.edu.tr

1. INTRODUCTION

Automatic Speech Recognition is basically the detection and recognition of human voice by an electronic device. This system converts speech to text using various detecting and pattern recognition methods. With the developments in computers and advanced processors, area of ASR has been sprouted. Air traffic control, ticket reservations, security and biometric identification, gaming, devices control in cars and home automation, robotic are some of these applications. Moreover, this system can be used to improve life quality of disabled and elderly persons.

The most advanced ASR systems are developed for popular and widely spoken languages, e.g. Chinese, Spanish and English. These languages have extensive transcriptional speech and text data to create acoustic and language models. Pronunciation dictionaries are generated with the help of pronunciation rules and acoustic units examined by linguists [1,2].

Turkish is an agglutinative language like Finnish or Hungarian. There are suffixes which are attached to word root. Therefore as the number of the different vocabularies gets higher, the performance of speech recognizers gets lower. Using a vocabulary of sub-words (morphemes) instead of words is also a common solution. Sub-word based ASR is commonly applied for languages such as Estonian, Finnish, Turkish, Thai, Hungarian, Slovenian and Czech, which are agglutinative languages [3]. A sub-word n-gram model can assign probabilities to unseen word forms [4]. In the past, a smaller vocabulary size was important for the recognition performance of Large Vocabulary Continuous Speech recognition (LVCSR) systems, but it is now useful for the latest Neural Network Language Models (NNLMs). Excessive vocabulary leads to increased size of the input and output layers. Therefore, it is not practical. Although class-based models can reduce the layers, methods or shortlists like hierarchical softmax, the sub-word models provide an effective and natural way to lower the dimensionality [5]. In ASR applications of agglutinative languages, Hungarian Finnish and Turkish, [3-8] sub-word based language models outperform word-based language models. Therefore sub-word based language model is preferred for training the proposed ASR system.

The GMM is generally used when calculating the probability (p (A | Q)) of the acoustic feature vectors (A) conditioned by the HMM phonetic / status tags (Q) in the HMM [9]. HMM is often used as acoustic model by the ASR systems. The phonemes in the utterance are predicted by using GMM-HMM model. Then the word or set of continuous words that were spoken are determined [10]. The posterior probability of each state in HMM is estimated by these architectures which use Perceptual Linear Predictions (PLPs) and Mel Frequency Cepstral Coefficient (MFCCs) as features. ASR problem for Turkish language has been tried to be solved by conventional methods. One of them is GMM-HMM which has been mostly used in the applications [11-13]. In [11], the Phoneme error Rate (PER) is 29.3%, the Word Error Rate (WER) observed in [12],[13] are 21.46% and 32.88% respectively. The performance of Turkish ASR can be improved by using advanced methods similar to Deep learning.

In ASR applications, Gaussian mixtures for speech recognition have successfully been replaced by the Deep Learning approach. Previous studies show that, in ASR applications, DNNs work better than GMM-HMM systems [10]. DNN includes trained hidden layers. Hidden layers provide large output which is required for big number of HMM states. Each phone is modelled by triphone HMMs which cause large number of states. Many studies show that DNN outperforms GMM at acoustic modelling which is used for ASRs having large datasets and vocabularies. Instead of using GMMs, DNNs are used to produce posterior probabilities over HMM states as output. In the study, pre-trained DBN-DNN is used when calculating the probability of the acoustic feature vectors. HMM or HMM states are used to model units of speech (phones, sub-phones, phone states, syllables, words, etc.). Deep Learning is still not effectively implemented in recognition of Turkish speech. There are few studies on DNN based ASR Applications of Turkish Language. In [9], an automatic dictation and keyword search system is designed for processing spoken lectures in Turkish. A DNN based LVCSR system is developed. The purposed system is trained by Turkish News TV program recordings and Law lecture video recordings. In [14], DNN is also used for building acoustic model. In [15], GMM and DNN based models are trained and tested using the corpus which is developed in [16]. The observed WERs are reported as 14.18%, 12.1% and 14.65% respectively. Considering the aforementioned studies the performance of Turkish ASR systems can be improved by applying sub-word (morpheme) based language models and replacing GMM acoustic modelling with DNN based models. All performance metrics are defined considering the sub-word and word recognition performances in the mentioned studies.

Sub-word (morpheme) based LM will be useful for improving the recognition rates of the agglutinative languages and prevents large vocabulary in the system. Root and suffixes are modelled separately in this model. In the proposed system, sub-word (morpheme) based language modelling is implemented. In addition to word, sub-word recognition, phoneme recognition is also supported. Each phoneme of Turkish language is also modelled as a sub-word. In this way, vocabulary coverage of the system has been enlarged.

Kaldi ASR toolbox, open speech recognition tools, is set up by Daniel Povey [17]. This tool can be used to implement DBN-HMM which builds language models and acoustic models. In the study, Kaldi handles the training and decoding for ASR of Turkish language.

The purpose of this work is suggesting a systematic approach for designing Deep Learning based ASR for Turkish language and improve the performance of the ASR by reducing effect on the PER and the WER. In this study, a standard corpus, “Turkish Microphone Speech v1.0” which was accepted by Linguistic Data Consortium (LDC) in 2005, is used to perform the suggested system. The corpus is used in previous studies [11],[12],[13],[18]. In the similar studies [36], [37], [19], the databases which are provided by authors are used and have not been used in other studies. The databases are not standard and not confirmed by LDC. Using the standard dataset confirms the accuracy of applied method and provides an opportunity to compare the study with the previous papers which are using the same dataset. For example, the result of GMM based recognition of our study is very similar to the result which is obtained in [11].

While the previous studies Sub-word based language model, is used for training. Due to Turkish is a phonemic language, all phonemes in Turkish are also trained as a sub-word and entered to the lexicon. In case, sub-word or word recognition fails, phoneme based recognition finds out the phoneme components of unrecognized word or sub-word. The phoneme components are concatenated and unrecognized word can be formed.

In [15] and [19], the Deep Learning based ASR systems support only word-based ASR and the performance of the proposed systems are measured only by Word Error Rate (WER) metric. In our study, the performance of the applied method is also measured by Phoneme Error Rate (PER) metric. Our previous studies [20],[21], on phoneme based ASR for Turkish Language was published previously in 2016 and 2017. Since Turkish is a phoneme-based language, out of vocabulary (OOV) words can be found out by the help of recognized phoneme components. In the literature review, any deep learning-based phoneme recognizer was not encountered for Turkish language.

For training and testing, DBN in which the hidden layers are pre-trained by features is used. Considering similar studies, standard (traditional) sequential DNN and Time Delay Neural Network (TDNN) structures, have been applied in references [15],[19]. As recommended in reference [15], applying DBN will be preferred for ASR applications. The results of our study show that, the accuracy of DBN is

1.5% higher than previous DNN and TDNN based studies.

The performance of DBN-HMM based recognizer in the proposed study is compared with the performance of GMM-HMM recognizer in [11] which uses the same corpus. PER is used for comparing the results. WER performance of the methods, GMM-HMM and DBN-HMM, are also compared.

The rest of the paper is organized as follows: In Section 2, Morphological structure of Turkish language is explained, Section 3 explains the language models which are used in ASR applications as well as the preferred model. The mostly used speech recognition methods, Deep Learning, DNN algorithm and codes are explained in Section 4. Experiments and results are described in Section 5. Section 6 and 7 presents the discussions on results and conclusions.

2. TURKISH MORPHOLOGY

Turkish has an agglutinative morphology. Addition of several suffixes can derive many new words from a single stem. Prefix is not used in Turkish. The examples below show concatenated verbal and nominal inflections. The nominal inflection is less complicated [22] (See Table 1).

Table 1. Verbal and Nominal inflection examples

Nominal Inflection	araba-m-da-ki-ler (*ones in my car*).
Verbal Inflection	koş-a-ma-dı-lar (*They could not run*)

In agglutinative languages, it is possible to add morpheme one after the other. Each morpheme carries some morphological information like tense, status, agreement etc. This property of agglutinative languages causes having large vocabulary including many words with the same stem but different endings. Since there are so many words in the dictionary, there will be a large number of OOV words which are not modelled [23]. Larger lexicon size degrades the speed of decoder in word-based speech recognition [24]. Due to high OOV, speech recognition methods for English applied to Turkish give low recognition results.

Another important property of Turkish in terms of language modelling is free word order. Subject-Object-Verb word order in Turkish is a typical characteristic, but other orders are possible under certain discourse conditions. The perplexity of N-gram language models increase, since a sentence can be constructed with different order of words without changing the meaning [24].

As a consequence, base-units different from words, sub-words (morphemes) must be utilized in Turkish LVCSR to solve coverage problem and large amount of training data is needed in order to reliably train language model parameters.

2.1. Phonemes in Turkish

Turkish is a phonemic language such as Japanese or Finnish that means in the written language each phoneme is symbolized by a letter. In other words, the written text and pronunciation match exactly. However, some consonants and vowels have different voices, depending on where they are produced in the sound path. [25].

During their formation, the voices of which the oral part of the breath channel is not blocked by the teeth, tongue, or lips are called vowels. In Turkish, vowels can be classified in three groups by jaw angle, shape of lips and position of tongue [26] (see Fig. 1).

Basic speech sounds in which the breath is blocked by the teeth, tongue, or lips and which can be combined with a vowel to form a syllable are called consonants. Consonants can be classified into two groups, which are voiced and unvoiced, by the state of vocal cords. Voiced consonants are consonant sounds which are made by vibrating the vocal chords. The vocal cords do not vibrate during the formation of unvoiced consonants. Consonants can also be classified by output types and output locations [26] (see Fig. 2).

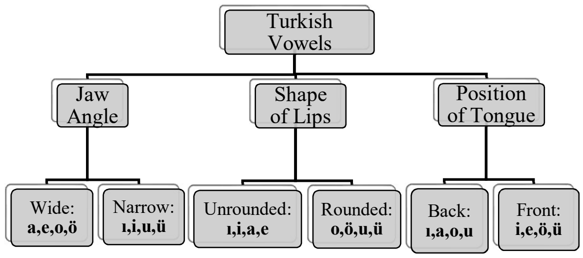

Figure 1. Turkish Vowel Classification

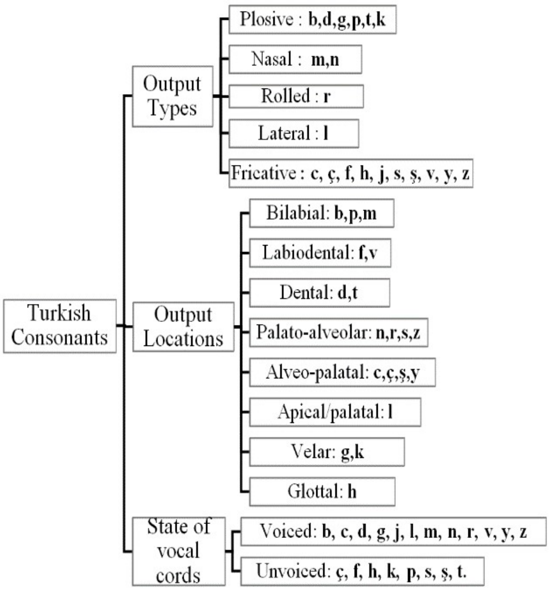

Figure 2. Turkish Consonant Classification

3. LANGUAGE MODEL (LM)

A Language Model (LM) obtains the probability of word existence based on text samples and develops the probabilistic models that can predict the following word in the chain of the given words. It contains a large list of words and their probability of occurrence. Larger models can predict sentences or paragraphs. Sub-word based model is preferred for the proposed system.

3.1. Word based LM

Word-based model is the most basic language modelling approach that uses words as recognition units. Words are the lexicon entries for speech recognition and language modelling and used as units for extracting probabilities from the training corpus [27] (see Fig. 3). The word-based LMs are preferred when modelling analytic and isolated languages which have a low morpheme-per-word ratio. For example, English and Mandarin Chinese are analytic languages.

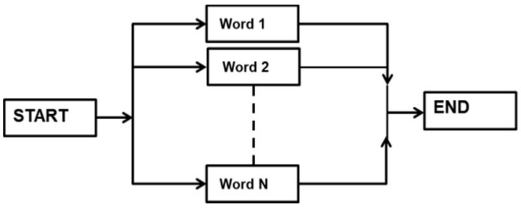

Figure 3. Word based LM

3.2. Sub-Word (Morpheme) based LM

Sub-word based LMs can improve the WER of agglutinative language ASR systems [3-8].

In Sub-word based LM, Morphemes are the basic units of the ASR system. (see Fig. 4). In Turkish language, A stem and following suffixes model the words according to the spelling rules and morphology. Bigram probabilities weight transitions between morphemes [7].

Phonetic rules are used for constructing the links between the stems and suffix lattices. The last phoneme of the stem and the last vowel determine the suffixes which may follow a particular stem. New stems can be added to the lattice automatically after this mapping is defined [29].

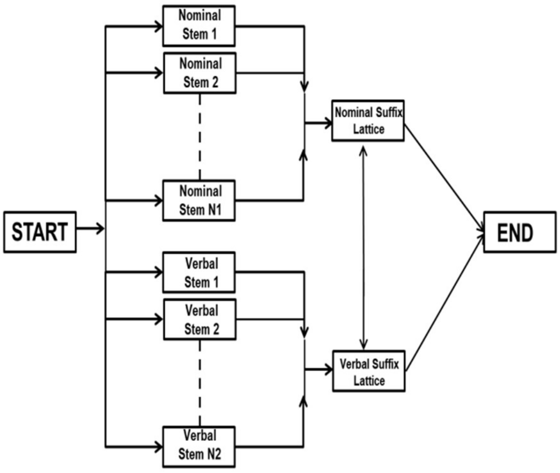

Figure 4. Sub-Word (Morpheme) based LM

4. METHODS OF SPEECH RECOGNITION

Statistical methods for ASR systems often use the HMM as an acoustic model. GMM is generally used when calculating the probability of the acoustic feature vectors for HMM. In last two decade, deep learning based speech recognition systems are developed for improving the recognition rate. Deep Learning approach has successfully replaced GMM for ASR. DNNs have been widely applied in acoustic model training and outperformed the statistical methods.

4.1. Statistical methods used in ASR

Mostly used and known ASR strategies are statistical based methods [30]. Blocks diagram of a statistical speech recognizer is shown in Fig. 5.

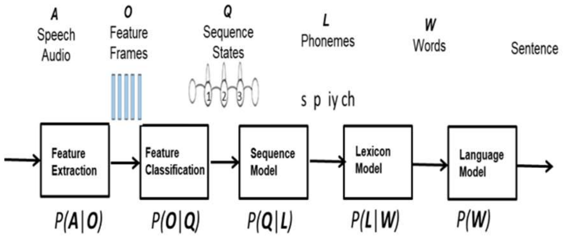

Figure 5. Statistical ASR system

The purpose of the ASR is converting speech audio to text. This procedure can be expressed statistically as below.

$O = (o_1, o_2, \dots, o_n)$ (Sequence of speech vectors, o_i, the vector at time i) is the set of acoustic observations and $W = (w_1, w_2, \dots, w_n)$ is the sequence of words. The maximum likelihood can be calculated as:

$$\widehat{W} = arg_w max P(W|O) = arg_w max \frac{P(W)P(O|W)}{P(O)} \quad (1)$$

Equation (1) uses Bayes rule and specifies the most likely word order. $P(O)$ represents the probability of the speech utterance. It is independent of the W sequence and can be neglected. Thus, (1) is simplified as shown below [30]:

$$\widehat{W} = arg_w max P(W)P(O|W) \quad (2)$$

Equation (2) contains two main factors:

The word sequence prior probability *P(W)*.

The probability of the acoustic data for the word sequence *P(O|W)*.

P(W) value depends on the LM and *P(O|W)* is calculated on the basis of the acoustic model. Both models can be generated separately, but they work together while recognizing a speech signal. The basis of the acoustical modelling is represented by HMMs.

4.1.1 Hidden Markov Models (HMM) for ASR

HMM method is known as state of art for speech recognition. GMM functions and state transition probabilities are used for classifying in HMM. HMMs are stochastic finite state machines and builds acoustic models and LMs in ASR applications [31]. They include a sequence of states related by transitions (see Fig. 6). The Markov process is named “hidden” because the observer cannot view the state sequence directly. A speech vector sequence which is generated from a Probability Density Function (PDF) for each state is observed.

As given in Fig. 6, A HMM assigns probabilities to a sequence of states. Consequently, the parameters below characterize an HMM: [30]

- a set of states $S = (s_1, s_2, \dots, s_N)$ State at time *t*:q_t
- States transition probabilities:
  $A = (a_{11}, a_{12}, \dots, a_{NN})$, each a_{ij} represents the transition probability from state *i* to state *j*;
- observation probabilities, each $B = b_i(o_t)$ defines the probability of an observation o_t which is built from the state i;
- the initial state distributions:

$$\pi = \{\pi_i = P[q_1 = s_i], i = 1, \dots, N\}. \quad (3)$$

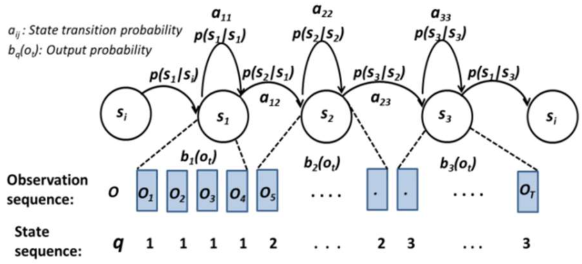

Figure 6. HMM representation

An HMM is indicated by *λ = (A, B, Π)*. HMMs suppose that the speech signal is stable in small time periods. HMMs are able to manage the speech signal variability well and they are good at speech modelling. In speech recognition models, speech is modeled by the states which are in order from left to right.

4.1.2 Gaussian Mixture Models (GMM)

Most widely used extension for standard HMMs is the mixture model of the state-output distributions. In the HMM Based Recognizer, Gauss distribution is used for modelling the state output distribution, assuming that the observed feature vectors are unimodal and symmetric [32].

In practice, speaker, gender and accent differences have tendency to create multiple modes in the data. Replacing the single Gaussian state-to-output distribution with a GMM which can model multi-modal and asymmetric data solves this problem (see Fig. 7).

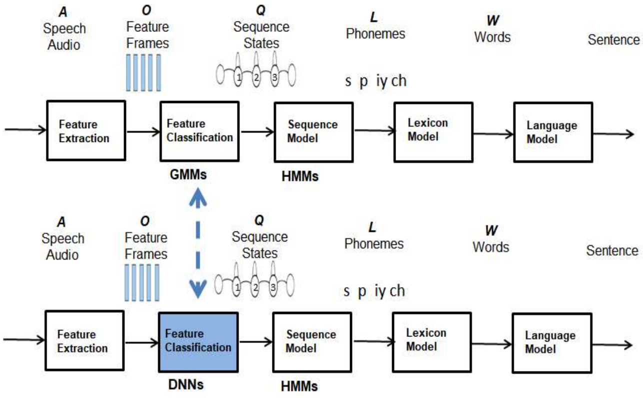

Figure 7. GMM-HMM and DNN-HMM based ASR systems

The GMM used for ASR, assumes that the spectral shape can be represented by a *M* component mixture model having parameters, mixture components $p_m(x)$ and component weights w_m. The general form of a univariate mixture model is:

$$p(x) = \sum_{m=1}^{M} w_m p_m(x) \tag{4}$$

where $p_m(x)$ are mixture components and w_m are weights. The mixture components in this case are Gaussian [33].

4.1.3 GMM and DNN in speech recognition:

The use of Gaussian mixtures in speech recognition applications has been replaced by the Deep Learning Approach. In ASR applications, DNNs work better than GMM based systems and they also left GMM's performance behind a large margin in some tasks [10] (see Fig. 7).

4.2. Deep Neural Networks (DNN)

Deep learning is an advanced architecture of ANN. More than one hidden layer is used for learning about the pattern and the features. Many complex signal patterns, for example, videos, images and speech, is successfully learned by means of many layers. These layers have nodes with nonlinear processing functions. Pre-training steps allow training large networks having millions of nodes.

DNNs with many hidden layers have been shown to perform better than GMMs in various speech recognition criteria. [34]. Hidden layers are used to detect patterns in data. More layers provide working on more complex data.

With nonlinear activation functions, DNN is able to model an arbitrary nonlinear function (projection from inputs to outputs). In each hidden unit, *j*, total input from the layer below, x_j, is mapped to the scalar state, y_j and sent to the layer above.

$$x_j = b_j + \sum_i y_i w_{ij} \quad (5) \qquad y_j = logistic(x_j) = \frac{1}{1 + e^{-x_j}} \quad (6)$$

b_j is the unit *j* bias, i is an index of units in the layer below, and w_{ij} is the weight for the link from i to j in the layer below.

For multiclass classification, the "softmax" linearity is used to get a class probability p_j

$$p_j = \frac{\exp(x_j)}{\sum_k \exp(x_k)} \tag{7}$$

where k is an index all over classes [34].

In the training phase of DNN networks, the derivative of the mismatch between the actual and target outputs is back-propagated. The initial weights can be set to small random values. Generatively, pre-training the DNN as a Deep Boltzmann Machine (DBM) or as a DBN is a better way of initialization then the enrolment samples are used to fine-tune [35].

4.2.1 Generative Pre-training:

In generative pre-training, feature detectors are initially designed for modelling the structure in the input data instead of designing feature detectors for discriminating between classes [34]. At first, one layer of feature detectors is trained by feature data then the trained feature detectors act as data for training the next layer. The multiple layers of feature detectors are prepared well for a discriminative “fine-tuning” phase by this generative "pre-training". During backpropagation, weights found in pre-training are slightly adjusted by DNN.

Restricted Boltzman Machine (RBM), a special Boltzman Machine, consists of a hidden unit layer and a visible unit layer having no hidden - hidden or visible-visible connections. The connections between hidden units and visible units are symmetrical and undirected. The value of hidden and visible units is often stochastic binary units. (1 or 0 depending on probability) [36].

The RBM structure is illustrated in Fig. 9. The model can be defined as follows for a set of values of (v, h) in an energy function[37]:

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\sum_{i\epsilon\text{visible}} a_i v_i - \sum_{j\epsilon\text{hidden}} b_j h_j - \sum_{i,j} v_i h_j w_{ij} \tag{8}$$

Where h_j and v_i are the binary states of visible unit i and hidden unit j respectively; w_{ij} is the weight between hj and $vi;$ b_j and a_i are the biases of h_j and v_i, respectively. The joint distribution over v and h is given as follows:

$$P(\mathbf{v}, \mathbf{h}; \theta) = \frac{1}{Y} \exp\left(-E(\mathbf{v}, \mathbf{h}; \theta)\right) \tag{9}$$

where Y is a partition function given as:

$$Y = \Sigma_v \Sigma_h e^{\{-E(v,h;\theta)\}} \tag{10}$$

Learning maximum likelihood using the contrastive divergence (CD) algorithm uses the training data for estimating the RBM model parameters θ [38].

In ASR applications, DBN-DNN produces posterior probability for each state of HMM according to the features. Speech units such as words, syllables, phones, phone states, sub-phones, etc. are modelled by HMM or states of HMM [39].

4.2.2 Deep Belief Networks (DBN):

DBN has made many important achievements in image and audio classification. The purpose of DBN is that learning the typical data features by training large amounts of unlabeled data in an unsupervised way. Stacked RBMs create a DBN [40].

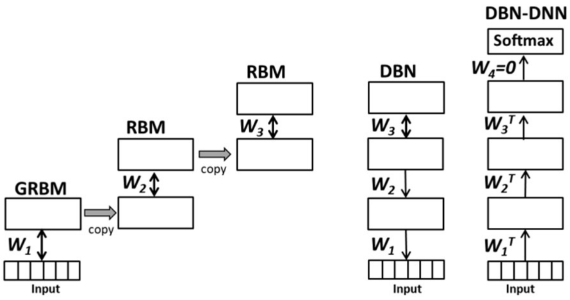

Figure 8. Deep Belief Networks [34]

The operations which are used for creating a DBN with three hidden layers and converting it to a pre-trained DBN-DNN are explained as follows (see Fig. 8). First a Gaussian-Bernoulli RBM (GRBM) is trained for modelling real-valued acoustic coefficient frames [34]. Training a RBM needs data which is the states of the binary hidden units of the GRBM. As many hidden layers as desired are created by repeating this process. The undirected connections of low-level RBMs are replaced with top-down, directed links. Therefore The RBM stack is transformed into a single generative model and forms a DBN. A “softmax” output layer containing each possible state of each HMM is added. Therefore, a pre-trained DBN-DNN is built. The DBN-DNN is then discriminatively trained. In this way, the HMM states of the input window are estimated.

Features of speech, MFCCs, are modelled by linear variables with Gaussian noise by giving a GRBM; the RBM energy function is given as:

$$E(\boldsymbol{v},\boldsymbol{h}) = \sum_{i\epsilon \text{vis}} \frac{(v_i-a_i)^2}{2\sigma_i^2} - \sum_{j\epsilon \text{hid}} b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij} \qquad (11)$$

Where b_j and a_i are the biases of h_j and v_i, respectively; and σ_i is the of the Gaussian noise standard deviation for visible unit i [34].

5. EXPERIMENTS AND ANALYSES

ASR application needs a Corpus which is a collection of texts of written or spoken language stored in a computer database. Corpora are the plural of corpus. Written texts in corpora are drawn from newspapers, books or magazines. Spoken corpora contain transcripts of spoken language.

The corpus “Turkish Microphone Speech” used in the experimental study is composed of 40 vocalized sentences from 120 speakers and 38 phonemes corresponding to 29 letters in Turkish alphabet [11].

In the experiments, GMM-HMM and DBN-HMM based two ASR systems are trained and tested. PER and WER are obtained for speaker independent situations.

In the purposed study, open source Kaldi codes are used for ASR algorithms. Voice recordings belonging to 120 people are used in the analyses. The speech data of 100 persons were used for the training and the data of 20 persons were used for the test.

5.1. Audio Corpus

The audio corpus, “Turkish Microphone Speech v1.0”, used in our study was collected from 193 speakers (104 male, 89 female) in the Department of Electrical and Electronics Engineering at Middle East Technical University (METU). A set of 40 sentences among 2462 sentences are selected randomly for each speaker. This is the method which is used to form TIMIT corpus. Each sentence is uttered once by the speakers.

The corpus named “Middle East Technical University Turkish Microphone Speech v.1.0” was accepted by Linguistic Data Consortium (LDC) in 2005 [11].

Turkish language has some non-Latin vowels such as Ü, Ö, İ and non-Latin consonants such as Ç, Ş, Ğ. These special letters are symbolized as follows (see Table 2) and it has been possible to add these special letters into Kaldi speech recognition toolbox.

Table 2. Special Turkish letters in Kaldi

Special Turkish Letter	Symbols in Kaldi
Ç	CH
Ğ	GH
İ	IY
Ö	OE
Ş	SH
Ü	UE

There have been studies using “Turkish Microphone Speech Corpus”, on Turkish speech recognition. These are [11], [12], [13] and [18]. In first study, the best phone recognition rate is 70.8% (PER=29.2%). In second study, word recognition rate is 78.54% (WER= 21.46%). In third study, word recognition rate is 67.12% (WER= 32.88%). In the fourth study, Common Vector Approach is used for classification. Vowel phoneme recognition rate is 48.75% and consonant phoneme recognition rate is 53.02%.

5.2. Kaldi ASR Toolbox

The Kaldi ASR toolbox is an open speech recognition tools set up by Daniel Povey. Various instructions of HMM Toolkit (HTK) are integrated and then DNN model is introduced. The system framework is shown in Fig. 9 [17].

Kaldi speech recognition toolkit performs training and decoding operations. A monophone system builds a triphone HMM for speech recognizer. The Triphone system uses the basis of alignments which are derived from the monophone system. Therefore, the alignments supplied by the GMM system is used for training the deep learning system [39]. Kaldi provides building acoustic models and LMs. Kaldi is written in C++ programming language and it is an open source speech recognition toolkit. The Kaldi toolkit includes several shell scripts and C++ executables. The codes are easy to understand, very modern and flexible. This is available on both Linux and Microsoft Windows operating systems [41].

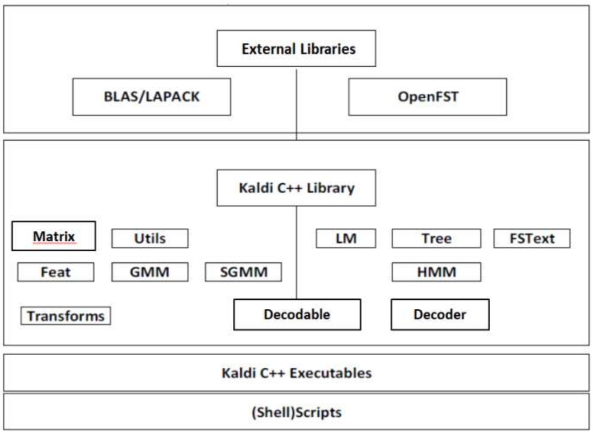

Figure 9. Kaldi system framework [17]

5.3. Data preparation (Pre-Processing)

Data preparation includes two main steps which are preparation of acoustic data and preparing language data.

Step1: Preparation of acoustic data: The acoustic data needs to be prepared before Kaldi ASR training. "Turkish Microphone Speech" corpus data which is used for training and testing are placed in the data folder (see Table 3). The data format must be converted to the desired form. Voice recordings belonging to 120 people are used in the analyses. The voice and sentence utterance of 100 people are placed in train folder. The data of 20 people for test is placed in test folder. In training 4000 (100p x 40) sentences voiced by 100 persons are used. 3416 words are used in the training stage. After audio data is prepared, the corresponding language and acoustic models are built. Then three files, wav.scp, text and utt2spk are created manually. These files also can be created by using written scripts.

Table 3. Contents of "data" folder

Files	Explanation
wav.scp	The paths of speaker audio files
Text	Speech content of speakers
utt2spk	Tells which utterance belongs to particular speaker
spk2gender	Speaker gender
Corpus	Every single utterance transcription in ASR system (40 sentences per speaker)

The MFCC features are extracted and stored in feats.scp file after running scripts. The files which are called feats.scp, wav.scp, utt2spk, text and spk2utt are placed in the data/test and data/train folder. Corpus is placed in data/local folder (see Table 3)

The feature vectors are obtained over 20 ms frames. These frames are represented by 13 parameters consisting of 12 MFCC parameters and frame energy. 50% overlapping between frames is applied. Attributes obtained from 7 consecutive frames are added one after the other to obtain 91 dimensions (13x7) feature vector representing 80 ms frame.

Step2: Preparing language data: A folder which is named "dict" is created in data/local directory. The files related with lexicon and phonemes are created in the folder (see Table 4).

Table 4. Contents of "dict" folder

Files	Explanation
lexicon.txt	phone translations of words
nonsilence_phones.txt	Phones in the language (39 phones in Turkish)
silence_phones.txt	lists silence phones

"Lexicon.txt" file contains every word in our dictionary with its phone translations. Some word translations to phones are shown in Table 7. "Nonsilence_phones.txt" file contains all of 38 phonemes in Turkish. “Silence_phones.txt” file lists silence phones.

5.4. Model Training

In Kaldi, shell and Perl scripts are used for decoding and training. DNN model is trained based on the HMM model by Kaldi [11].

For Language Modelling SRILM Toolkit for Kaldi is used. Training and Decoding is made in three steps:

Step1: GMM-HMM model based on three phonemes (triphone) is trained using data in train folder and it is decoded by data in test folder.

Step 2: DBN-HMM model based on triphone is trained and decoded.

DBN is formed by a stack of six RBMs (see Fig. 10). In DBN-HMM, sigmoid activation function is used in hidden units having dimension of 1024 (see Table 5).

Table 5. DBN-DNN Parameters

Number of RBMs	Hidden Layer	Learning rates
6	Sigmoid dim:1024	Initial: $8x10^{-3}$ Final: $7.81x10^{-6}$

Optimal initial learning rate for sigmoid in similar applications is 8x10-3. Final learning rate is 7.81x10-6 when training ends (see Table 5).

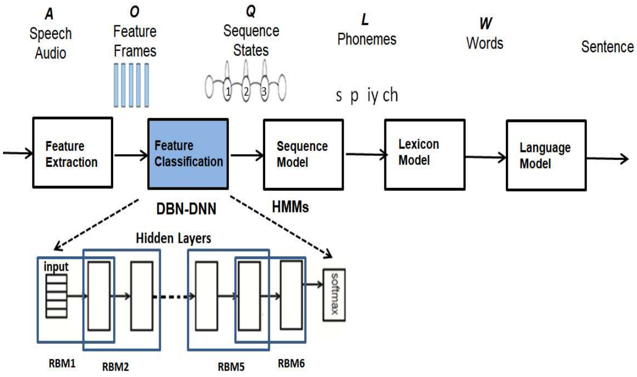

Figure 10. System Architecture

Cross-validation technique is applied for making DBN-HMM model robust against new data which are out of the training set. By this way overfitting is prevented. The dataset is split into separate training and test subset. At every iteration, the test data subset is changed and the system is retrained. After every iteration, error function is measured. If the amount of error change is low enough, Early Stopping is applied and training process is finalized at this point.

After training and decoding process we got the following PER values. The PER of GMM-HMM, and DBN-HMM recognizers which are built on Kaldi is also given (see Table 6).

Table 6. Phoneme Error rates (PERs)

Model	PER
GMM-HMM	30.64%
DBN-HMM	24.80%

The GMM-HMM PER value which we obtain is consistent with the value of the previous study [11] in which the same corpus used.

Secondly, word recognition performance has been tested. The phoneme components of corpus vocabulary, stems and morphemes are given to the system as input. The phoneme components of some words and endings are on Table 7. The WER of the classifiers are given (see Table 8).

Table 7. Some stems, suffixes, words in corpus and their phoneme translation

Stem/Ending	Phoneme Components - Lexicon
SÖZLEŞME	S OE Z L EE SH M EE
NİN	NN IY NN
YAP	Y AA P
ILMASINI	I LL M AA S I NN I
KİM	K IY M
İSTİYOR	IY S T IY Y O RH
BUNDA	B U NN D AA
KÖŞE	K OE SH EE
GÖRÜN	G OE RR UE NN
NÜR	NN UE RH
ŞEKİL	SH EE K IY L
DE	D EE
KES	K E S
İLME	IY L M EE
MİŞ	M IY S

Table 8. Word Error rates (WERs)

Model	WER
GMM-HMM	17.21%
DBN-HMM	13.04%

6. DICUSSION OF RESULTS

In this study DBN-HMM based ASR system for Turkish language has been developed. The proposed system has been compared with GMM-HMM based ASR recognition method. In suggested system, GMM has been replaced with Deep Learning architecture. The recognition performance of the proposed method has been compared with the GMM-HMM method applied in [11]. “Turkish Microphone Speech Corpus v1.0” accepted by LDC is used for training and testing. Voice recordings belonging to 120 people are used in the analyses. The voice recording records of 100 persons were used for the training and 20 of them were used for the test.

In [11], GMM-HMM speech recognizer, trained by same corpus is included. The PER is measured as 29.3%. In the proposed study, the PER of Kaldi speech recognizer using GMM-HMM architecture is measured as 30.64%. Here we can see the results are similar and consistent.

When we train DBN-HMM speech recognizer with the same corpus, we get a degreased PER of 24.8% and improved recognition performance.

Considering word recognizer, The WER of GMM-HMM is measured as 17.21%. In the suggested Deep Learning based structure, the WER is 13.04% which is lower than in GMM-HMM.

In the similar study [15], both GMM-HMM and DNN-HMM systems are trained by the database which is recorded and prepared by authors. The WER of systems are measured as 17.40% and 14.65% respectively. The obtained results are similar to the presented study. The differences between two studies are the used speech corpus and the applied Deep Learning method. DBN is used instead of Feed Forward Neural Network in our study. As recommended in [15], applying DBN will be preferred for ASR applications. The results of our study show that, using DBN has approximately 1.5% higher performance effect on the system. Unlike the study [15], the corpus “Turkish Microphone Speech v1.0” which is a standard dataset and accepted by LDC, has been used in the proposed study. Using an approved dataset which can be accessed by other researchers, improves accessibility, comparability and evaluability of published studies using the same database.

7. CONCLUSIONS

This paper proposed a DBN based Turkish phoneme and speech recognizer for improving the quality and performance of ASR for Turkish as an agglutinative language. The proposed system recognized words in the system vocabulary and phoneme components of OOV words. Sub-word (morpheme) LM which is commonly used for agglutinative languages is preferred. Each phoneme of Turkish language is modelled as a sub-word. Considering that Turkish language is a phoneme-based language, it is desired to reach an approximate result by concatenating the letter components of the words that cannot be recognized or are not in the dictionary. The performance of the suggested DBN based system is compared with conventional recognition method, GMM based HMM using the same dataset. In both architectures, HMM is used for sequence and language modelling. Application and comparison results show that the performance of the recognition has improved considering PER and WER criteria. Moreover, from these results it is observed that the Deep Learning models can produce deeper and more accurate feature probabilities and that the speech recognizer has a more discriminating ability. In future studies Long Short-Term Memory and Recurrent Neural Networks Language modelling with the same corpus can be performed and language-specific solutions for ASR in Turkish language can be developed.

Research and Publication Ethics

This paper has been prepared within the scope of international research and publication ethics.

Ethics Committee Approval

This paper does not require any ethics committee permission or special permission.

Conflict of Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this paper.

REFERENCES

[1] 'The most spoken languages worldwide (native speakers in millions) - Statistica, The Statistics portal', https:// www.statista.com / statistics / 266808 / the- most-spoken-languages-worldwide/, accessed 19 November 2018.

[2] Wang L, Tomg R, Leung C, Sivadas S, Ni C, Ma, B., 'Cloud-Based Automatic Speech recognition System for Southeast Asian Languages', International Conference on Orange Technologies (ICOT), IEEE, 2017, pp. 147-150.

[3] Varjokallio, M., Kurimo, M., Virpioja, S., 'Learning a Subword Vocabulary Based on Unigram Likelihood', IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2013, pp. 7-12.

[4] Varjokallio, M., Kurimo, 'A Word – Level Token – Passing Decoder for Subword N-gram LVCSR', IEEE Spoken Language Technology Workshop (SLT), 2014, pp. 495-500.

[5] Smit, P., Gangireddy, S., R., Enarvi, S., Virpioja, S., Kurimo, M., 'Character-Based Units for Unlimited Vocabulary Continuous Speech Recognition', IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2017, pp. 149-156.

[6] Mihajlik, P., Tüske, Z., Tárjan, B., Németh, B., Fegyó, T., 'Improved Recognition of Spontaneous Hungarian Speech-Morphological and Acoustic Modeling Techniques for a Less Resourced Task', IEEE Transactions On Audio, Speech, And Language Processing, Vol. 18, No. 6, August 2010, pp. 1588-1600

[7] Arısoy E., Saraclar M., 'Language Modelling Approaches for Turkish Large Vocabulary Continuous Speech Recognition Based on Lattice Rescoring', 14th Signal Processing and Communications Applications, IEEE, 2006

[8] Aksungurlu T., Parlak S., Sak H, Saraçlar M., 'Comparison of Language Modelling Approaches for Turkish Broadcast News', 16th Signal Processing, Communication and Applications Conference, IEEE, 2008

[9] Arısoy, E., 'Devoloping an Automatic Transcription and Retrieval system for Spoken Lectures in Turkish', 25th Signal Processing and Communications Applications Conference (SIU), IEEE, 2017

[10] Dhankar, A., 'Study of deep Learning and CMU Sphinx in Automatic Speech Recognition', International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017, pp. 2296-2301.

[11] Salor,O., Pellom, B.L., Çiloğlu, T., Demirekler, M., 'Turkish speech corpora and recognition tools developed by porting SONIC: (Towards multilingual speech recognition)', Computer Speech and Language, Elsevier, 2007, 21, pp. 580–593.

[12] Bayer, A., O., Çiloglu, T., Yondem, M., T., 'Investigation of Different Language Models for Turkish Speech Recognition', 14th Signal Processing and Communications Applications, IEEE, 2006

[13] Susman, D., Köprü, S., Yazıcı, A., 'Turkish Large Vocabulary Continuous Speech Recognition By Using Limited Audio Corpus', 20th Signal Processing and Communications Applications Conference (SIU), IEEE, 2012

[14] Arısoy E., Saraclar M., 'Compositional Neural Network Language Models for Agglutinative Languages', Interspeech 2016, San Francisco, USA, pp. 3494-3498

[15] Büyük, O., Kimanuka, U. A., 'Turkish Speech Recognition Based on Deep Neural Networks', Süleyman Demirel University Journal of Natural and Applied Sciences Volume 22, Special Issue, 2018, pp. 319-329

[16] Büyük, O., 'A new database for Turkish speech recognition on mobile devices and initial speech recognition results using the database', Pamukkale University Journal of Engineering Sciences Volume 24-2, 2018, pp. 180-184

[17] Ruan, W., Gan, Z., Liu, B., Guo Y., 'An Improved Tibetan Lhasa Speech Recognition Method Based on Deep Neural Network', 10th International Conference on Intelligent Computation Technology and Automation, IEEE, 2017, pp. 303-306

[18] Keser, S., Edizkan, R., 'Phoneme-Based Isolated Turkish Word Recognition With Subspace Classifier', 17th Signal Processing and Communications Applications Conference , IEEE, 2009.

[19] Asefisaray, B., Haznedaroğlu , A., Erden, M., Arslan, L., M., "Transfer Learning for Automatic Speech Recognition Systems", 26th Signal Processing and Communications Applications Conference (SIU), 2018

[20] Tombaloğlu, B., Erdem, H., "Development of a MFCC-SVM Based Turkish Speech Recognition system", 24th Signal Processing and Communication Application Conference (SIU), 2016

[21] Tombaloğlu, B., Erdem, H., " A SVM based speech to text converter for Turkish language", 25th Signal Processing and Communication Application Conference (SIU), 2017

[22] Arısoy E., Saraclar M., 'Lattice Extension and Vocabulary Adaptation for Turkish LVCSR', IEEE Transactıons on Audio, Speech and Language Processıng, vol. 17, no. 1, 2009

[23] Tunalı, V., 'A Speaker Dependent Large Vocabulary Isolated Word Speech Recognition System for Turkish', Msc. Thesis, Marmara University, 2005.

[24] Büyük O., 'Sub-Word Language Modelling for Turkish Speech Recognition', Msc. Thesis, Sabanci University, 2005.

[25] Salor, Ö., Pellom,B., Çiloğlu, T., Hacıoğlu, K. and Demirekler, M., 'On developing new text and audio corpora and speech recognition tools for the Turkish language, ICSLP-2002: Inter. Conf. On Spoken Language Processing, Denver, Colorado USA, pp. 349–352.

[26] Ergenç, İ., 'Konuşma Dili ve Türkçenin söyleyiş sözlüğü', Multilingual, Istanbul, 2002, p. 486.

[27] Arısoy E., Saraclar M., 'Turkish Dictation System for Broadcast News Applications', 13th European Signal Processing Conference, 2005.

[28] Arısoy E., Dutagacı H., Saraclar M., 'A unified language model for large vocabulary continuous speech recognition of Turkish', Signal Processing 86 , Elsevier, 2006, pp. 2844-2862.

[29] Dutagacı H, 'Statistical Language Models for Large Vocabulary Turkish Speech Recognition', Msc. Thesis, Boğaziçi University, 2002.

[30] Schiopu, D., 'Using Statistical Methods in a Speech Recognition System for Romanian Language', 12th IFAC Conference on Programmable Devices and Embedded Systems, 25-27 September 2013, Velke Karlovice, Czech Republic, pp. 99-103.

[31] Köklükaya, E, Coşkun, İ., "Endüktif Öğrenmeyi Kullanarak Konuşmayı Tanıma". Sakarya University Journal of Science 7, 2003, pp. 87-94.

[32] Gales, M., Young, S., 'The Application of Hidden Markov Models in Speech Recognition', Foundations and Trends in Signal Processing, Vol. 1, No. 3, 2007, pp. 195–304.

[33] Stuttle, M., N., ‘A Gaussian Mixture Model Spectral Representation for Speech Recognition’, Ph.D. Thesis, Cambridge University, 2003.

[34] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B., ‘Deep Neural Networks for Acoustic Modelling in Speech Recognition’, IEEE Signal Processing Magazine,Volume: 29 , Issue: 6 , Nov. 2012, pp. 82-97

[35] Alam, M. R., Bennamoun M., Togneri R., Sohel F., ‘Deep Neural Networks for Mobile Person Recognition with Audio-Visual Signals’, Mobile Biometrics, 2017, p. 97-129.

[36] Banumathi, A., C., Chandra, Dr. E., ‘Deep Learning Architectures, Algorithms for Speech Recognition: An Overview’, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 7, Issue 1, January 2017, pp. 213-220.

[37] Siniscalchi, S., M., Svendsen, T., Lee, C., 'An artificial neural network approach to automatic speech processing', Neurocomputing, Elsevier, 2014, Vol. 140, pp. 326-338.

[38] Sharan, R. V., Moir, T. J., `An overview of applications and advancements in automatic sound recognition`, Neurocomputing, Elsevier, 2016, Vol. 200, pp. 22-34.

[39] Sustika, R., Yuliani, A. R., Zaenudin, E., Pardede, H. F., `On Comparison of Deep Learning Architectures for Distant Speech Recognition', 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), IEEE, 2017.

[40] Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E., `A survey of deep neural network architectures and their applications', Neurocomputing, Elsevier, 2017, Vol. 234, pp. 533-541.

[41] Yadava, G T., Jayanna, H S., `Creating Language and Acoustic Models using Kaldi to Build An Automatic Speech Recognition System for Kannada Language', 2nd IEEE International Conference On Recent Trends in Electronics Information and Communication Technology (RTEICT), May 19-20, 2017, India, IEEE, pp. 161-165