Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Review Article

Turk J Elec Eng & Comp Sci
(2020) 28: 3253 – 3269
© TÜBİTAK
doi:10.3906/elk-2001-38

A detailed survey of Turkish automatic speech recognition

Recep Sinan ARSLAN*, Necaattin BARIŞÇI
Department of Computer Engineering, Faculty of Technology, Gazi University, Ankara, Turkey

Received: 07.01.2020 • **Accepted/Published Online:** 10.07.2020 • **Final Version:** 30.11.2020

Abstract: Significant improvements have been made in automatic speech recognition (ASR) systems in terms of both the general technology and the software used. Despite these advancements, however, there is still an important difference between the recognition performance of humans and machines. This work focuses on the studies conducted in the field of Turkish speech recognition, the progress made in such studies in recent years, the language-specific constraints, the performance results achieved in the applications developed to date, and the development of a general scheme for researchers wishing to develop an ASR system for the Turkish language. A comprehensive study on the Turkish language, including all literature and speech recognition systems, was prepared. There has been significant progress in the languages of developed countries in terms of ASR systems, but it was observed in this work that the opposite is true for the Turkish language. The lack of labor force and material resources remains the biggest obstacle to developing better systems. As a result, concrete studies are needed to achieve an ASR system with performance equivalent to the human level for the Turkish language.

Key words: Turkish speech processing, automatic speech recognition (ASR), feature extraction, classifiers, end-point detection, robust ASR

1. Introduction

Speech is the most effective method of communication among people. For this reason, it is the most natural approach for data exchange, and the demand for people to interact vocally with computers is steadily increasing. Much work has been done to meet this demand of people, such as performing simple tasks with human-machine interaction or simulating people's speaking abilities or speech to text applications [1]. In recent years, the progress needed for speech recognition systems has been accomplished by studies carried out in this field and the studies for the development of more robust systems have been increasing in number.

Speech recognition systems mainly include the modeling of signals and the matching of the patterns. The modeling of signals is the process of converting voice signals to a number of parameters and it includes 4 process steps: spectral shaping, feature extraction, parameter transformation, and statistical modeling. Spectral shaping is the process of converting a speech signal to a digital signal and emphasizing the important frequency parts in the signal. Feature extraction is the process of obtaining different features from speech signals such as power, pitch, and vocal tract configuration. Parameter transformation is the conversion of these features into signal parameters. Statistical modeling involves the conversion of parameters into observation vectors [2–4].

Research on speech recognition models started in 1936 at Bell Labs. The first ideas about speech recognition were related to the evaluation of the acoustic phonetic structure. In 1952, a speaker-dependent isolated digit recognition system was developed by Davis et al. in Bell Labs [5]. In early 1960, special hardware

*Correspondence: recep.sinan.arslan@gazi.edu.tr was produced to accelerate computers sufficiently, which recognized vowels [6]. Usable structures were proposed by Velichko and Zagoruyko [7], Sakoe and Chiba [8], and Itakura [9] regarding isolated recognition in the 1970s. Studies carried out until the 1980s were focused on the isolation of speech. Problems with continuous speech recognition became the subject of research for the first time in the 1980s. In those years, there was a transition from the template-based approach used for statistical modeling methods. A hidden Markov model structure started to be used in speech recognition applications [10, 11]. Bayesian computation and clustering techniques were also proposed in the 1990s. The aim was to find maximum probability values and to produce more successful results [12]. The adaptive learning problem was solved after the 2000s and learning-based automatic speech recognition (ASR) systems were developed [13]. Today, more successful results are obtained with the development and spread of GPU architecture and the use of deep learning systems [14]. In recent years, many postgraduate theses related to the Turkish language have been completed. An end-to-end Turkish speech recognition system was developed in the doctoral thesis of Asefisaray [15] in 2019. The aim was to predict and correct punctuation marks in the text. Ultimately, 63.4% of commas, 68.9% of periods, and 82.5% of question marks were corrected. Environmental speech recognition was proposed by Arslan [16] in 2018. Classification was performed with 99.9% accuracy in the detection of gunshot sounds and 98.4% in the detection of screams and vehicular accident sounds. Finally, in the master's thesis prepared by Akın [17], it was aimed to develop a model based on biometric recognition through discrimination of twin voices, which were difficult to separate visually and auditorily. In the tests performed with a test set containing 78 twin voices, 91% of them were identified correctly. Although there has been considerable progress in ASR systems over the last 60 years, there are still many problems that need to be solved and many issues to be improved [18].

Automatic speech recognition has started to be used in many areas, such as problems requiring human/machine interaction, travel information management, weather reports, automatic data entering, speech recognition, and data access, because ASR models produce more successful results. In the modeling of speech recognition systems, the workflow given in Figure 1 is usually followed. Accordingly, after the audio signal is received in the system via the microphone, the preprocessing stage is started. At this stage, the extraction of speech samples, digitizing, use of various filters when necessary, labeling, and transformation into a specified format are carried out. Here, the aim is to produce a state of speech that is less simple, and free of speech variations and noise. In the next step, parameters are obtained from sampled audio signals. Various calculations are performed to extract the properties of the speech at certain time intervals. The data obtained in this step are critical values for speech recognition and they provide critical keys. In the next step, a comparison of the predicted words from the parameters and important data is performed with the language models. Accordingly, the data set of words is used when guessing words from the acoustic model output. Thus, it is aimed to choose the most correct word [18].

This paper presents a detailed analysis of the scientific studies carried out in the field of Turkish speech recognition applications, focusing on the performance values that were achieved in sample projects the goals that remain unmet, and the projections for work that might be conducted in the future. Information on the characteristics of the Turkish language and relevant studies is given in Section 2, while the results of previous studies and the gaps in the literature are addressed in Section 3.

2. Turkish speech recognition techniques

In this section, techniques used in Turkish speech recognition (TSR) systems are summarized. The subjects of the study and the algorithms used for classification are categorized based on audio signal analysis methods and

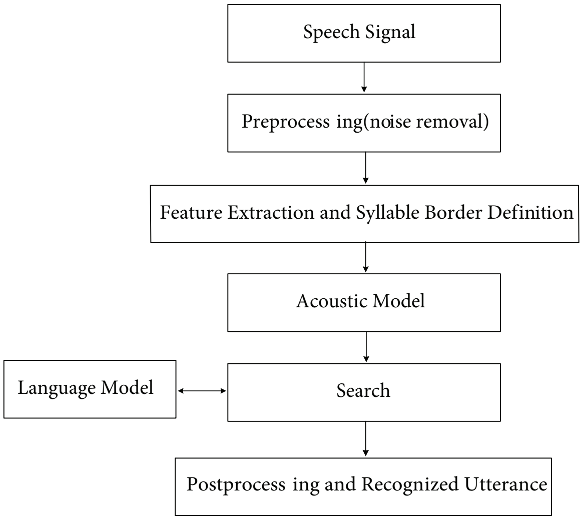
Speech Signal
Preprocess ing(noise removal)
Feature Extraction and Syllable Border Definition
Acoustic Model
Language Model
Search
Postprocess ing and Recognized Utterance

Figure. Basic flow of speech recognition systems [18].

all studies carried out on TSR systems are presented.

2.1. Brief overview of the Turkish language

The Turkish language, spoken in Turkey, belongs to the Oghuz group of the Turkic languages within the Ural-Altaic language family [19–21]. It is spoken in a wide geographical area mainly spanning Turkey, Cyprus, Iraq, the Balkans, Middle Asia, and Central European countries. It is an agglutinative language. The alphabet has 29 letters [22]. Today, the development of voice, communication, and processing technologies necessitates the development of robust and reliable speech recognition engines that take into consideration the vocabulary, acoustic parameters, communication structures, and language models of the Turkic languages, which are considered as the languages of the Silk Road and are spoken by about 100 million people [23]. The development of speech recognition systems is a challenging process. Different speech recognition software must be created for each language. Therefore, the speech recognition software developed for each language has to solve the problems arising out of that language's specific features [24]. Since the Turkish language is an agglutinative language, the number of words that can be derived, compared to other languages, cannot be easily delimited. More than 100,000 morphological variants of a verb stem alone can be generated with derivational affixes and inflectional suffixes. When this derivation process is carried out for each stem in Turkish, billions of different morphological structures emerge [25]. Other morphologically rich languages, such as Czech, Finnish, and Hungarian, experience a similar problem in terms of speech recognition [26]. This feature of the language increases the number of out-of-vocabulary words and requires working with large audio sources. Language models used in speech recognition systems are expected to have a low number of nondictionary words in order for them to contribute to performance [27]. This phenomenon causes the development of TSR systems to be a challenging problem because it affects the recognizer performance directly and causes a decrease. Moreover, the fact that the Turkish language allows free word alignment makes it difficult to create robust language models [28]. Alternative suggestions have been put forward to solve these problems. In the context of the recognition unit, the use of various subword units as an alternative to words was proposed [29, 30]. Previous studies carried out in the field of TSR are generally based on the language model, and they are word-based, stem-affix-based, syllable-based, and morphology-based. The results obtained vary based on the n-gram level and the source of the data set used [28]. Even though there are many papers in the literature on speech recognition, very few of those studies are on the Turkish language. The majority of studies are on languages such as English, Chinese, Spanish, German, and French [23]. In development of intelligent devices, speech recognition techniques for many foreign languages, especially English, are largely used, and new technological devices come with speech recognition support, but none have Turkish support [31]. One of the first studies carried out in the field of speech recognition for the Turkish language was about isolated recognition, conducted in 1999 [32]. Subsequent studies have been based on new units that are alternatives to words (letters, roots, root + suffix, stem, etc.) for large vocabulary systems. Mengusoglu and Deroo were the first to adapt the stems and stem-affixes in Turkish as speech recognition units [33]. In the last decade, extensive studies have been carried out on TSR in the field of automatic writing and retrieval of news programs [34].

2.2. Types of speech recognition

Speech recognition systems are classified based on the speech recognition approaches, speech utterance, speaker model, feature extraction, and terminology [4, 35].

2.2.1. Speech utterance

Speech recognition systems are basically divided into four categories based on speech utterance [35]. Isolated word recognition requires remaining silent on both sides of the sample windows. This does not mean that it always accepts a single word; it requires a single simultaneous expression. This type is good for command-based systems but is not suitable for multirow word input systems. Word borders are certain and the expressions are pronounced clearly. The disadvantage of this type is that it is affected if different borders are determined. Connected word recognition is a system that expects small pauses between words during speech. Although it is like an isolated word recognition system, it allows processing with insignificant interruptions between expressions. Continuous speech recognition deals with speech in which words are connected together instead of being separated by silence. Unknown border information on words, coarticulation, the production of the surrounding phonemes, and speech speed affect the performance of continuous speech recognition systems. Finally, spontaneous speech recognition, which is actually the system expected by the users, is a system of recognition of natural or unprepared speech such as interviews, discussions, dialogues etc. Such ASR systems, which are able to recognize voices spontaneously, recognize various natural sounds (hmm, uh, silence, etc.) [35].

2.2.2. Speaker model

People, pronounce words differently due to their different physical characteristics. In the process of developing speech recognition systems, they are divided into 2 classes as the main consideration. For speaker-dependent models, the speech recognition system can only work with specific speakers. Such systems require less work, yield results that are more accurate, and solve problems more easily. Speaker-independent models, on the other hand, are systems that support speaker diversity. They use data sets where different speech samples are collected. The development of such systems is cumbersome, requires difficult problems to be solved, and may yield results with less precision (accuracy). However, these systems are more flexible and can cope with different situations [35].

2.2.3. Based on volume of words

The vocabulary with which a speech recognition system works is an element that affects the unpredictability of the system, the resolution of prerequisites by the system, and the accuracy of recognition. While some applications can work with a few words (command-based), others may need stronger word data sets. It is possible to classify sets of words as small (up to ten words), medium (up to 100 words), large (up to thousands of words) and very large [36]. When all available studies are analyzed in light of the sample of studies given in Table 1, it is understood that no specific type was preferred in the studies carried out for Turkish and that these studies performed modeling based on different voice types depending on the data set used for the problems that the researchers wanted to create solutions for (conversion from voice to text, voice command systems, voice-activated security systems, automatic response centers). It was confirmed that higher performance values were achieved in field-based studies or in studies where the volume of words was smaller, and better classification results were obtained in isolated and speaker-independent models.

Table 1. List of some studies in Turkish according to speech recognition types.

Name of authors	Utterance	Speaker model	Volume of words	Classification unit	Performance
Buyuk et al. [37]	Continuous	Speaker-independent	Word set on radiology	Word and sentence based	SV* :90 % LV*: 44 %
Peker [31]	Isolated	Speaker-independent	Numbers in the range 0–9	Command-based	84 %
Caner and Üstün [38]	Isolated	Speaker-dependent	Turkish vowels	Word-based (vowels, consonants)	70 %
Keser and Edizkan [39]	Isolated	Speaker-independent	METU 1.0 Dataset	Phoneme	70 %–80 %
Büyük et al. [40]	Continuous	Speaker-independent	2151 Test Data	Syllable-based	3 % increase performance

2.3. Feature extraction methodologies

Feature extraction techniques are utilized to extract speech signals for use in speech recognition systems. The aim is to maintain the distinctive power of the signals while reducing the size of the input vector. With regard to the problem of the classification process in speech recognition systems, the number of training and test vectors grows with the size of the input given, so the feature of the speech signal needs to be extracted. There are methods commonly used to extract features from speech signal samples [41]. Some of them are analogue to digital transformation, short-term frequency analysis, discrete Fourier transformation, filter bank analysis, autocorrelation analysis, cepstral processing, linear prediction analysis, and wavelet transformation. In this section, some feature extraction methods used in Turkish language studies are summarized.

2.3.1. Mel frequency cepstral coefficients (MFCC)

This is one of the feature extraction methods most commonly used in speech recognition. It is not a linear structure in terms of detecting the frequency content of the sounds for speech signals [42]. It possesses features that increase the accuracy of recognition process, including delta and double delta values. Moreover, it lacks some features in the design of robust speech recognition systems in noisy speech [43]. The MFCC feature extraction process has several steps. The preemphasis stage is used to increase the energy level at high frequencies of input signals. Thus, the data in these regions are taken as better and these data are useful in hidden Markov model(HMM) training. Windowing slices the input signal into discrete time segments. This is performed by using a window of N milliseconds in width and offsets of M milliseconds in length. Discrete Fourier transform(DFT) is applied to each window. This results in the magnitude and phase representation of the signal. DFT spectrum results include information for each frequency. However, humans are less sensitive to frequencies above 1 kHz. For this reason, the logarithmic mel scale is applied to DFT outputs. Finally, cosine transformation is applied to the mel spectrum data and mel-cepstral coefficients are obtained. Thus, the MFCC features are reached [44].

2.3.2. Linear predictive coding (LPC)

This is one of the most powerful signal analysis techniques. It is used to estimate the basic parameters of speech. It provides accurate estimation of speech parameters and is also an effective speech modeling method. The fundamental idea of the LPC method is that a speech sample can be estimated as a linear combination of past speech examples. With the minimization of the differences between real speech and predicted speech, unique parameters or prediction coefficients can be identified. These coefficients constitute the basis of the LPC method. These estimated coefficients are used to generate the parameters of robust speech recognition systems known as cepstral coefficients [45]. The first step in LPC processing starts with digitization. The preemphasis step is applied for spectral smoothing of the speech signal and to make it less sensitive to external influences. Then, in the second step, the preemphasized speech signal is blocked into frames of *‘n’* samples, with adjacent frames being separated by *‘m’*samples. In the next step, windowing is performed to minimize signal discontinuity at the beginning and end of each frame. Then for each frame of the windowed signal, the highest correlation value is obtained by automatic correlation. Finally, the autocorrelated value of each frame is converted to LPC parameters [46]. The studies and performance results using different feature extraction methods are shown in Table 2.

Table 2. Summary of previous studies (extraction methods).

Name of author	Feature extraction	Details of study	Performance
Palaz et al. [23]	MFCC	Is a speaker-independent speech recognition application. Classification was made using HMM model for continuous and isolated speeches.	Without noise: 96.76% With noise: 65.61%
Gelegin and Bolat [47]	MFCC	Is a command-based (20), speaker-dependent speech recognition application. Classification was made through HMM.	Results ranging between 92% and 100%.
Yakar and Aşlıyan [48]	LPC + MFCC	Is a syllable-based, isolated Turkish speech recognition application. Worked with a 200-word vocabulary. Classification was made through HMM.	20% increase was achieved
Oflazoglu and Yıldırım [49]	MFCC + LPC	Study was conducted in determining the feeling angry – not angry with the help of the acoustic feature. Classification was made through naïve Bayes which is a sort of classifier.	75.8%
Akın et al. [50]	PLP	Solution was provided for nonvocabulary problems for agglutinative languages such as Turkish	20.01

2.4. Speech recognition approaches

2.4.1. Template based approaches

In template-based approaches, for the recognition of unknown speech, the best match needs to be found. This match is achieved by comparing recognition outputs with prerecorded template words [51]. It has highly accurate classification performance in word-based recognition systems [47, 52–54]. The basic idea of this technique is to create a collection of sample speech patterns and keep them as a reference while catching the most suitable match by comparing the unknown expressions to these patterns. It allows eliminating many segmentation and classification problems that arise from studies conducted with subunits as the units to be recognized do not necessarily have to be broken down into words or subword units. It is possible to perform the recognition process on sentences. On the other hand, if a speech recognition system with a large vocabulary is developed, it becomes a rather expensive and impractical method, as a large number of word templates will be needed. It is a preferable method in modeling with a fixed number of templates [55]. In the study carried out by Edizkan et al. [52] in 2007, it was aimed to control the direction of navigational vehicles using 5 command templates. Using the MFCC method, the features of the commands were extracted and sample templates were generated. A 100% success rate was achieved in tests performed using MATLAB for this application, which is speaker-dependent. In another study [53] carried out using 6 command templates, a recognition rate of 78% was reached, and an average of 100% was accomplished in an other study [54] using 60 sentence templates. In studies conducted for the Turkish language, the general characteristics of the template-based approach were learned and implemented and satisfying values were achieved in recognition performance.

2.4.2. Probabilistic approaches

It is more appropriate to use probabilistic models when it is necessary to work with incomplete data because of uncertainties or insufficient data sets. The presence of sounds that could be mixed with each other, speaker variability, presence of contextual effects, and words with similar audio characteristics are crucial problems for speech recognition systems [56, 57]. The HMM is the probabilistic approach most commonly used in TSR applications. Compared to template-based approaches, HMM models can be applied to more general problems and have a good mathematical basis [18]. In approximately 50% of the studies conducted on the Turkish language, HMM was used. The performance of the recognition varies depending on the feature extraction technique, type of problem, and vocabulary used. The studies with the best results are shown in Table 3.

2.4.3. Artificial neural networks (ANNs)

The ANN approach is a method aimed at enabling computers to perform the processes of displaying, analyzing, and characterizing speech based on the set of acoustic features extracted. The methods in this category use pragmatic information for phonetic, syntactic and semantic features and even segmentation and labeling, aiming to learn the relationships between phonetic events. The focus of this approach is on the representation of information and the integrated use of information sources [18]. In a study carried out by Özbey and Bayar [25], a TSR model was developed. The CMU Sphinx voice recognition library was used. With classification using an ANN, 70% accuracy was achieved. In another study, a syllable-based TSR system was developed [61]. In the continuous speech recognition system, developed using a time delay neural network (TDNN), classification was performed with accuracy of 65.6%. In another study [62] in which an ANN was used, work was carried out on the integration of biometric systems and security systems. A 99% classification rate was achieved following 200 test results.

Table 3. Summary of previous studies (stochastic approach).

Name of authors	Vocabulary	Details of study	Percentage of error
Aksoylar et al. [24]	Metu 1.0, SUVoice	Models with limited and larger vocabulary were compared. It runs on phoneme basis.	Recognition of numbers: 1% Names of places: 1.2% Sports news: 37%
Türk and Arslan [58]	14 Phonemes, 125 words	A phoneme and word-based isolated model was generated. Designed to be used in speech therapy.	Speaker-independent phoneme based: 15.1%
Arısoy and Saraçlar [59]	Large word set	Design of a speech recognition system with a large vocabulary	It runs with less than 20% error.
Arısoy and Arslan [60]	Morph, Root + Post, Root models	A system was designed for Turkish paper news to be automatically dictated.	54%

2.4.4. Support vector machines (SVMs)

The use of SVMs is a learning method applied to solve classification and regression problems based on statistical learning theory and systems with the least structural risk. It generates a solution by turning these problems into second-degree programming problems with the help of local solutions and this is what makes the SVM method most advantageous compared to other techniques. It is highly competent in making generalizations (i.e. defining missing information with the help of generalization). It is also used in many different fields [63, 64]. In a study carried out by Eray et al. [65] in 2018, a TSR system was developed using an SVM. Twenty words commonly used in Turkish computer systems were selected. Tests conducted using soft margin (SM-SVM) and least square (LS-SVM) methods yielded success rates of 91% and 71%, respectively. In another study conducted in 2016, a phoneme-based practice was performed with a performance rate of 84% [66]. In addition, in a study in which Bayesian networks and J48 classifiers [67] were compared with the SVM method, it was demonstrated that the best results for the detection of emotions were achieved with the SVM and classification was performed with 80% accuracy. In a 2016 study in which music genre classification was conducted for Turkish music, variance in the error rate was shown when different features were used in decision support systems [68]. Tombaloğlu and Erdem [69] conducted a study for the Turkish language, in which a phoneme-based MFCC feature extraction algorithm was used and classification was done using decision support systems. A performance value of 84% was obtained in speaker-dependent modeling and 51% in speaker-independent systems. It was seen that decision support systems were preferred in the studies conducted in the Turkish language in 2018, as well. They are most likely preferred as they are suitable for systems with low resources, such as those for the Turkish language, and they are able to achieve better values in classification performance.

2.4.5. Deep learning

Deep learning is a machine learning technique where information processing stages, consisting of several layers in hierarchical architectures, are used for structure classification and feature or representation learning [70]. It allows computational models consisting of multiple process layers to learn the representation of data with multiple abstraction layers. These methods are used in many areas and particularly in visual object recognition, object detection, genetics, and speech recognition. To demonstrate how the internal parameters of a machine used to calculate its display in each layer from its presentation at previous layers need to change, deep learning discovers complex structures in large data sets using the back propagation algorithm [71]. In order to shorten the duration of neural network training, distributed architecture was proposed [72–75]. However, conventional neural network approaches have become more popular recently due to the fact that GPU architecture is included in deep learning processes with a reduction in training time. Therefore, the use of large amounts of data shared on the internet in the speech recognition field has become widespread as access to powerful systematic infrastructures is growing easier [76]. A deep learning model was preferred in applications such as Cortana, Xbox, Skype Translator, Google Now, Siri, Baidu, and Nuance, which are commercially available and developed by companies such as Google, Microsoft, and Apple [77]. When the studies carried out in the Turkish language using deep learning were analyzed, it was noticed that only two such studies have been carried out. In the study conducted by Arısoy in 2018 [34], work was performed on the automatic typing of lectures in the Turkish language. Classification was performed with a language model of 200,000 words and deep learning was used, and an error rate of 14.2% was achieved. The study has the best error rate among studies on the Turkish language conducted with large vocabularies. The second such study, carried out by Arslan and Barışçı in 2018 [78], evaluated the contributions of different optimization techniques to classification performance for deep learning. The results showed that the gradient descent, proximal gradient descent and RMSPROP algorithms were better than others for Turkish.

2.5. Effects of design of language model on Turkish speech recognition systems

Continuous speech recognition with a large vocabulary can be performed with relatively high success rates for languages that do not have rich morphological features, such as English. However, this rate is quite low for agglutinative languages such as Turkish. The main reason for this is that language models created for other languages are insufficient for agglutinative languages [79]. In agglutinative languages, many words can be derived from a word root with the help of affixes. For this reason, it is theoretically not possible to create a language model that would cover all the words for a language such as Turkish. In order to find a solution to this problem, different approaches and model structures such as phoneme-based, affix + roots, body + affixes, and morphology-based approaches were proposed in place of using words as recognition units [80]. In a study conducted with a morphology-based approach [81], the word-based model yielded an error rate of 38.8%, while a morphology-based model yielded an error rate of 33.9%. The contribution of various data amounts to the recognition performance for languages with a large number of nonvocabulary words was analyzed and a performance increase of 2% was achieved when the data size was increased from 25 million to 200 million [82]. The extent to which the increased size of compilation would contribute to the performance was evaluated by Behnam et al. and it was confirmed that creating a large amount of language models would directly contribute to the performance, but the expected performance increase could not be achieved after a certain level [27]. In another study [83], in which voice recognition performance was evaluated where tri-grams were used in the design of the language model instead of bi-grams, it was stated that a significant 12% increase was achieved in recognition performance in tri-gram cases. In tests performed with 220 words, while the recognition rate was 72.99% with bi-gram models, it was 85.12% in tri-gram modeling. In the study by Arısoy et al. [79], they evaluated the contribution to recognition performance if words were chosen as the recognition unit, vocabulary was increased, the scope was widened, and the language model was selected to be a more complex one. Compared to other word-based systems, a better performance was achieved. Consequently, a speaker-independent model could be designed with an error rate of 15%. Similarly, in studies [60, 84] conducted by the same individuals, the aim was to dictate the outputs of the Turkish news system. With that aim, instead of the standard word or subword units, different word units such as root, suffix, morphology, and lexicon, which are smaller than words, were used in the recognition system. In this way, efforts were made to reduce the number of non-vocabulary words faced in agglutinative languages. In order to solve the problems arising from the structure of the language in the design of speech recognition systems in Turkish, different proposals were made in the design of the language model as given above and the results were evaluated. End-to-end speech recognition systems [85, 86], where the language model is never used to solve the problem, were proposed.This is thought to be a particularly applicable approach to the solution of this problem, which is a significant issue for the Turkish language. Thus, it will be possible to develop robust Turkish speech recognition systems with rich acoustic data.

2.6. Speech recognition datasets in Turkish

In the process of developing speech recognition systems for the Turkish language, there is a need for data sets to produce both acoustic models and language models. Thus, it is possible to carry out the training of models in a quick manner and to perform studies aimed at getting better results. Audio data sets have a wide scope of applications in ASR systems. In addition, they are used in many fields such as automatic voice synthesis, coding, speaker recognition [87, 88], and language recognition and verification. Speaker variety is very important in speech data sets and has a direct impact on recognition performance. This variety is due to variable speech speed, changing emotions and mental complexity and ambient noise. Voice recording can be performed in silent studios or in noisy environments. These recordings may be from television and radio broadcasts or telephone calls. Application-specific databases can also be created [18]. The text data sets that can be used to create language models are shown in Table 4 and audio data sets that can be used in the production of acoustic models for the Turkish language are shown in Table 5. The distribution of 5 audio data sets prepared between 2006 and 2019, given in Table 4, was made through the Linguistic Data Consortium (LDC)[1]. In general, it includes phone call records and “Voice of America”[2] news recordings. These are spontaneous speech data. They are all sets distributed for payment and are available in Turkish speech recognition applications. In Table 5, there are databases, prepared for the Turkish language containing only text data, and which can be used for language modeling. Except for the study published in 1994, these data are up-to-date, all published after 2015. In total, these sets contain large amounts of text data.

3. Discussion and conclusion

Speech has significant potential for communications to happen and for interaction with computers. Today, the most important activity in the field of speech recognition is to get the voice and turn it into text and commands. In this paper, a systemic report on Turkish speech recognition studies, their results, and the

[1] Linguistic Data Consortium (2020). Home Page [online]. Website https://www.ldc.upenn.edu/ [accessed 01 May 2018].

[2] Voice of America (2020). ????? [online]. Website https://www.amerikaninsesi.com/ [accessed 01 June 2018].

[3] Linguistic Data Consortium (2020). TS Wikipedia Corpus [online]. Website https://catalog.ldc.upenn.edu/docs/LDC2015T15/TS Wikipedia Documentation [accessed 03 September 2018].

[4] Linguistic Data Consortium (2020). ECI Multilingual Text [online]. Website https://catalog.ldc.upenn.edu/docs/LDC94T5/ [accessed 07 September 2018].

[5] Linguistic Data Consortium (2020). Appen ButterHill Language Specific Peculiarities Document for Turkish as Spoken in Turkey [online]. Website https://catalog.ldc.upenn.edu/docs/LDC2016S10/LSPfinal.pdf [accessed 09 August 2018].

Table 4. Data sets that can be used in generating language models.

Name of datasets	Number of units	Recording environment	Transcription based on
2007 CoNLL Shared Task – Basque, Catalan, Czech and Turkish (updated) [89]	3.1 GB	Text books, news, literature studies	Sentences
CoNLL Shared Task-Ten Languages-METU-Sabancı Turkish Dependency Treebank [90, 91]	3.1 GB	Text books, news, literature studies	Sentences
TS Wikipedia[3]	1.6 million web pages	Wikipedia	Speech tags, morphological analysis, lemmas, bi-grams and tri-grams
ECI Multilingual Text[4]	283K	Articles published in papers and magazines dictionaries and news	Sentences

Table 5. Data sets that can be used in generating acoustic models.

Name of dataset	Number of Units	Speech style	Transcription based on
Multilanguage conversational telephone speech 2011—Turkish [92]	18.6 h	Conversation spontaneous	Sentences
NIST language recognition evaluation test set [93]	215 h (23 languages)	Spontaneous	Sentences (word-based)
Turkish broadcast news speech and transcripts [94]	130 min speech record	Spontaneous	Sentences (word-based)
IARPA Babel Turkish language Pack IARPA-babel105b-v.0.5[5]	213 h	Conversation spontaneous	Sentences
Sentimental conversation database in Turkish [95]	5304 video recordings	Spontaneous	Sentences
Middle East Technical University Microphone Speech v.1.0 [96]	2462 sentence, 500 min	Reading	Sentences (word-based)

techniques applied has been presented along with a summary of the Turkish language. A general overview of the specific methods has been presented for better handling of different situations faced by ASR systems, and speech analysis and acoustic modeling aspects have mainly been discussed. In addition, feature extraction methods, and classification algorithms have been described. The potential areas of the application of research carried out in this field and the resources and tools developed by individual researchers have been shown, as well. The main components of the Turkish speech recognition systems examined as well as the techniques used in the relevant studies in this field have been reviewed and the details of some case studies were given. In addition, future possible studies and recommendations for the Turkish language have been presented. The aim here is to help researchers with the present and future situations of Turkish speech recognition procedures. When the recent developments regarding speech recognition systems in the Turkish language were reviewed, it was seen that the combination of HMM, MFCC, and SVM or MFCC and ANN modeling was used together in the studies carried out. The accuracy levels of the systems in which MFCC was used were higher and they yielded better performances compared to other methods. Moreover, it was shown that there are language-specific problems to be solved in order to develop speech recognition applications in Turkish and these problems have had significant adverse effects on recognition performance. An increase in the language resources for the Turkish language and the diversification of the data sets would allow studies in this field to expand. There is also a large amount of unprocessed data available from Turkish web pages, news sources, and blogs. To use these resources, it is necessary to develop complex techniques and tools. In doing so, it would be possible to develop systems that are more robust. Moreover, sharing the researches in this area would be very useful for fast and cooperative projects. Although Turkish speech recognition draws the attention of researchers, there are still many efforts remaining to be made to make this field more mature. The Turkish language has a historical literary past, a rich heritage, and importance for a large population, but systematic and scientific studies to create systems that work and convert speech to text have not yet been completed. Comprehensive and usable speech recognition software has not yet been made available for users. The studies conducted so far are focused on general speech recognition and are in the very early stages of development. This review has shown that there is still a long way to go for developing a realistic Turkish speech recognition system.

References

[1] Tran DT. Fuzzy approaches to speech and speaker recognition. PhD, Canberra University, Canberra, Australia, 2000.

[2] Bilmes J. EE516 Computer Speech Processing: Lecture: 2. Washington, USA: University of Washington Department of Electrical Engineering Press, 2015.

[3] Picone JW. Signal modelling techniques in speech recognition. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP); Minneapolis, USA; 1993. pp. 1215-1247.

[4] Unnibhavi AH, Jangamshetti DS. A survey of speech recognition on South Indian languages. In: International Conference on Signal Processing, Communication, Power and Embedded Systems; Odisha, India; 2016. pp. 1122-1126.

[5] Davis KH, Biddulph R, Balashek S. Automatic recognition of spoken digits. Journal of the Acoustic Society of America 1952; 24 (6): 637-642. doi: 10.1121/1.1906946

[6] Suzuki J, Nakata K. Recognition of Japanese vowels preliminary to the recognition of speech. Journal of the Radio Research Laboratory 1961; 37 (8): 193-212.

[7] Velichko VM, Zagoruyko NG. Automatic recognition of 200 words. International Journal of Man-Machine 1970; 2 (3): 223-234. doi: 10.1016/S0020-7373(70)80008-6

[8] Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transaction on Acoustics, Speech, Signal Processing 1978; 26 (1): 43-49. doi: 10.1109/TASSP.1978.1163055

[9] Itakura F. Minimum prediction residula applied to speech recognition. IEEE Transactions on Acoustics, Speech, Signal Processing 1975; 23 (1): 67-72.

[10] Rabiner LR, Juang BH. An introduction to hidden Markov models. IEEE ASSP Magazine 1986; (3) 1: 4-16. doi: 10.1109/MASSP.1986.1165342

[11] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of IEEE 1989; 77 (2): 257-286. doi: 10.1109/5.18626

[12] Juang BH, Furui S. Automatic speech recognition and understanding: A first step toward natural human machine communication. Proceedings of IEEE 2000; 88 (8): 1142-1165. doi: 10.1109/5.880077

[13] Riccardi G, Hakkani-Tur D. Active learning: theory and applications to automatic speech recognition. IEEE Transactions On Speech And Audio Processing 2005; 13 (4): 504-511. doi: 10.1109/TSA.2005.848882

[14] Hinton G, Deng L, Dong Y, George E, Jaitly N et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Processing Magazine 2012; 29 (6): 82-97.

[15] Asefisaray B. End-to-end speech recognition model: experiments in Turkish. PhD, Hacettepe University, Ankara, Turkey, 2018.

[16] Arslan Y. Detection and recognition of sounds from hazardous events for surveillance applications. PhD, Yıldırım Beyazıt University, Ankara, Turkey, 2018.

[17] Akın C. Voice recognition system with score level fusion methods and embedded system design. MSc, İstanbul Technical University, İstanbul, Turkey, 2019.

[18] Anusuya MA, Katti SK. Speech recognition by machine: a review. International Journal of Computer Science and Information Security 2009; 4 (3): 181-205.

[19] Tekin T. Altaic Languages, The Encyclopedia of Languages and Linguistic. Brno, Czech Republic: Masaryk University Press, 1994.

[20] Starostin SA, Dybo AV, Mudrak OA. An Etymological Dictionary of Altaic Languages. Leiden, Netherlands: Brill Publishers, 2003.

[21] Poppe N. Introduction to Altaic Linguistics. Wiesbaden, Germany: Ural-Altaische Bibliotek, 1965.

[22] Uzun, NE. Türkçenin dünya dilleri arasındaki yeri üzerine. Ankara Üniversitesi Dil ve Tarih-Coğrafya Fakültesi Türkoloji Dergisi 2012; 19 (2): 115-134.

[23] Palaz H, Kanak A, Bicil Y, Doğan MU, İslam T. TREN- Turkish speech recognition platform. In: IEEE Signal Processing and Communications Applications Conference; Kayseri, Turkey; 2005. pp. 1-4.

[24] Aksoylar C, Mutluergil SO, Erdogan H. The anatomy of a Turkish speech recognition system. In: IEEE Signal Processing and Communications Applications Conference; Antalya, Turkey; 2009. pp. 512-515.

[25] Özbey C., Bayar S. Automatic speech recognition: generating and testing generic acoustic model for Turkish. In: 19. Akademik Bilişim Konferansı; Aksaray, Turkey; 2017. pp. 1-6.

[26] Arısoy E. Statistical and discriminative language modelling for Turkish large vocabulary continuous speech recognition. PhD, Boğaziçi University, İstanbul, Turkey, 2009.

[27] Asefisaray B, Mengüşoğlu E, Hacıömeroğlu M, Sever H. How does language model size effects speech recognition accuracy for the Turkish language? Pamukkale University Journal of Engineering Sciences 2016; 22 (2): 100-105.

[28] Susman D, Köprü S, Yazıcı A. Turkish large vocabulary continuous speech recognition by using limited audio corpus. In: IEEE Signal Processing and Communications Applications Conference; Muğla, Turkey; 2012. pp. 1-4.

[29] Erdoğan H, Büyük O, Oflazer K. Incorporating language constraints in sub-word based speech recognition. In: IEEE Signal Processing and Communications Applications Conference; Kayseri, Turkey; 2005. pp. 1-6.

[30] Sak H, Saraçlar M, Güngör T. Morphology-based and sub-word language modelling for Turkish speech recognition. In: IEEE Signal Processing and Communications Applications Conference; Diyarbakır, Turkey; 2010. pp. 5402-5405.

[31] Peker O. Speech recognition by Turkish word. In: Signal Processing and Communication Application Conference (SIU); Zonguldak, Turkey; 2016. pp. 1737-1740.

[32] Arslan LM. Digit recognition application of Turkish continuous speech recognition (In Turkish). In: IEEE Signal Processing and Communications Applications Conference; Antalya, Turkey; 2009. pp. 64-67.

[33] Mengüşoğlu E, Deroo O. Turkish LVCSR: database preparation and language modelling for an agglutinative language. In: International Conference on Acoustics, Speech and Signal Processing; Salt-Lake City, USA; 2001. pp. 4018-4023.

[34] Arısoy E. Developing an automatic transcription and retrieval system for spoken lectures in Turkish. In: IEEE Signal Processing and Communications Applications Conference; Antalya, Turkey; 2017. pp. 1-4.

[35] Vanajakshi P, Mathivanan MA. Detailed survey on large vocabulary continuous speech recognition techniques. In: International Conference on Computer Communication and Informatics; Coimbatore, India; 2017. pp. 1-7.

[36] Yalçın N. Speech recognition theory and techniques. Kastamonu Educational Journal 2008; 16 (1): 249-266.

[37] Buyuk O, Haznedaroglu A, Arslan LM. Turkish speech recognition software with adaptable language model. In: IEEE Signal Processing and Communications Applications Conference; Eskişehir, Turkey; 2007. pp. 1-4.

[38] Caner M, Üstün SV. An application of speaker recognition using artificial neural networks. Journal of Engineering Sciences 2006; 12 (2): 279-284.

[39] Keser S, Edizkan R. Phonem-based isolated Turkish word recognition with subspace classifier. In: IEEE Signal Processing and Communications Applications Conference; Antalya, Turkey; 2009. pp. 93-96.

[40] Büyük O, Erdoğan H, Oflazer K. Using hybrid lexicon units and incorporating language constraints in speech recognition. In: IEEE Signal Processing and Communications Applications Conference; Kayseri, Turkey; 2005. pp. 111-114.

[41] Yadav M, Alam MA. Speech recognition: a review. International Journal of Research in Electronics and Computer Engineering (IJRECE) 2018; 6: 1-9.

[42] Vibha T. MFCC and its applications in speaker recognition. International Journal on Emerging Technologies 2010; 1 (1): 19-22.

[43] Chandresekar P, Chapaneri S, Jayaswal D. Automatic speech emotion recognition: a survey. In: International Conference on Circuits, Systems, Communication and Information Technology Applications; Mumbai, India; 2014. pp. 341-346.

[44] Karpagavalli S, Chandra EA. Review on automatic speech recognition architecture and approaches. International Journal of Signal Processing, Image Processing and Pattern Recognition 2016; 9 (4): 393-404. doi: 10.14257/ijsip.2016.9.4.34

[45] Madan A, Gupta D. Speech feature extraction and classification a comparative review. International Journal of Computer Applications 2014; 90 (9): 20-25. doi: 10.5120/15603-4392

[46] Thasleem TM, Narayanan NK. A robust approach for Malayalam cv speech unit recognition using artificial neural network. In: Proceedings of the International Conference on Image Processing Computer Vision and Pattern Recognition; Las Vegas, NV, USA; 2011. pp. 52-56.

[47] Gelegin İ, Bolat B. Computer control with a discrete-word speech recognition system. In: Elektrik-Elektronik ve Bilgisayar Sempozyumu; Elazığ, Turkey; 2011. pp. 1-4 (in Turkish).

[48] Yakar Ö, Aşlıyan R. Turkish speech recognition using hidden Markov model. In: Akademik Bilişim Konferansı; Aydın, Turkey; 2016. pp. 1-7.

[49] Oflazoğlu Ç, Yıldırım S. Anger recognition in Turkish speech using acoustic information. In: IEEE Signal Processing and Communications Applications Conference; Muğla, Turkey; 2012. pp. 1-4.

[50] Akın AA, Demir C, Doğan MU. Improving sub-word language modelling for Turkish speech recognition. In: IEEE Signal Processing and Communications Applications Conference; Muğla, Turkey; 2012. pp. 1-4.

[51] Myers CS, Rabiner LR. A level building dynamic time warping algorithm for connected word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing 1981; 29 (2): 284-297. doi: 10.1109/TASSP.1981.1163527

[52] Edizkan R, Tiryaki B, Büyükcan T, Uzun İ. Mobile vehicle control with voice command recognition. In: 9. Akademik Bilişim Konferansı; Kütahya, Turkey; 2007. pp. 1-10.

[53] Baygin M, Karaköse M. Real time voice recognition based smart home application. In: IEEE Signal Processing and Communications Applications Conference; Muğla, Turkey; 2012. pp. 1-4.

[54] Yalçın N, Bay ÖF. HTK ile konuşmacıdan bağımsız Türkçe konuşma tanıma sistemi oluşturma. Gazi Üniversitesi Endüstriyel Sanatlar Eğitim Fakültesi Dergisi 2007; 21: 79-97.

[55] Gaikwad SK, Gawali BW, Yannawar P. A review on speech recognition technique. International Journal of Computer Applications 2010; 10 (3): 16-24. doi: 10.5120/1462-1976

[56] Varga AP, Moore RK. Hidden Markov model decomposition of speech and noise. In: International Conference on Acoustics, Speech and Signal Processing; Albuquerque, Mexico; 1990. pp. 845-848.

[57] Weintraub M, Murveit H, Cohen M, Price P, Bernstein J. Linguistic constraints in hidden Markov model based speech recognition. In: International Conference on Acoustics, Speech and Signal Processing; Glasgow, Scotland; 1989. pp. 699-702.

[58] Türk O, Arslan LM. Speech recognition methods for speech therapy. In: IEEE Signal Processing and Communications Applications Conference (SIU); Kuşadası, Turkey; 2004. pp. 410-413.

[59] Arısoy E, Saraçlar M. Speech recognition for Turkish broadcast news. In: IEEE Signal Processing and Communications Applications Conference (SIU); Eskişehir, Turkey; 2007. pp. 1-4.

[60] Arısoy E, Arslan LM. Turkish dictation system for Broadcast news applications. In: IEEE Signal Processing and Communications Applications Conference (SIU); Kayseri, Turkey; 2005. pp. 629-632.

[61] Can B, Artuner H. A syllable-based Turkish speech recognition system by using time delay neural networks. In: International Conference on Soft Computing and Pattern Recognition (SoCPaR); Hanoi, Vietnam; 2013. pp. 219-224.

[62] Dede G, Sazlı MH. Analysis of biometric systems from pattern recognition perspective and voice recognition module simulation. In: Elektrik, Elektronik ve Bilgisayar Mühendisliği Ulusal Kongresi; İstanbul, Turkey; 2010. pp. 1-5.

[63] Cristianini N, Taylor JS. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge, England: Cambridge University Press, 2000.

[64] Schölkopf B, Burges CJC, Smola AJ. Advances in Kernel Methods: Support Vector Learning. Cambridge, England: The MIT Press, 1999.

[65] Eray O, Tokat S, İplikci S. An application of speech recognition with support vector machines. In: International Symposium on Digital Forensic and Security (ISDFS); Antalya, Turkey; 2018. pp. 1-6.

[66] Tombaloğlu B, Erdem H. Development of a MFCC-SVM based Turkish speech recognition system. In: IEEE Signal Processing and Communications Applications Conference (SIU); Zonguldak, Turkey; 2016. pp. 929-932.

[67] Oflazoğlu Ç, Yıldırım S. Binary classification performances of emotion classes for Turkish emotional speech. In: IEEE Signal Processing and Communications Applications Conference (SIU); Malatya, Turkey; 2015. pp. 2353-2356.

[68] Çoban Ö, Özyer GT. Music genre classification from Turkish lyrics. In: IEEE Signal Processing and Communication Application Conference (SIU); Zonguldak, Turkey; 2016. pp. 101-104.

[69] Tombaloğlu B, Erdem H. A SVM based speech to text converter for Turkish language. In: IEEE Signal Processing and Communications Applications Conference (SIU); Antalya, Turkey; 2017. pp. 1-4.

[70] Wan J, Wang D, Hoi S, Wu P, Zhu J et al. Deep learning for content-based image retrieval: a comprehensive study. In: Proceedings of the 22nd ACM International Conference on Multimedia; Orlando, FL, USA; 2014. pp. 157-166.

[71] LeCun Y, Bengio Y, Hinton G. Deep learning: review. Nature 2015; 521: 436-444. doi: 10.1038/nature14539

[72] Uzun İ, Edizkan R. Performance improvement of distributed Turkish continuous speech recognition system in case of burst packet loses. In: IEEE Signal Processing, Communication and Application Conference; Aydın, Turkey; 2008. pp. 1-4.

[73] Uzun İ, Edizkan R. Isolated speech recognition over internet protocol in distributed systems. In: IEEE Signal Processing, Communication and Application Conference (SIU); Aydın, Turkey; 2008. pp. 1-4.

[74] Saraçlar M, Arısoy E, Uzun İ, Edizkan R. Analysis and compensation of sparse packet losses in distributed Turkish continuous speech recognition system. In: IEEE Signal Processing, Communication and Application Conference (SIU); Aydın, Turkey; 2008. pp. 1-4.

[75] Uzun İ, Edizkan R. Performance improvement in distributed Turkish continuous speech recognition system using packet loss concealment techniques. In: Innovations in Intelligent Systems and Applications; İstanbul, Turkey; 2011. pp. 375-378.

[76] Anwer A, Mahmood O. Breast cancer diagnosis using deep learning methods (In Turkish). MSc, University of Turkish Aeronautical Association Institute of Science and Technology, Ankara, Turkey, 2017.

[77] Li J, Deng L, Haeb UR, Gong Y. Robust automatic speech recognition: a bridge to practical applications. Netherlands: Academic Press, 2015.

[78] Arslan RS, Barışçı N. The effect of different optimization techniques on end-to-end Turkish speech recognition systems that use connectionist temporal classification. In: International Symposium on Multidisciplinary Studies and Innovative Technologies; Ankara, Turkey; 2018. pp. 1-6.

[79] Arısoy E, Dutağacı H, Arslan LM. A unified model for large vocabulary continuous speech recognition of Turkish. Signal Processing 2006; 86 (10): 2844-2862. doi: 10.1016/j.sigpro.2005.12.002

[80] Bayer AO, Çiloğlu T, Yöndem MT. Investigation of different language models for Turkish speech recognition. In: IEEE Signal Processing and Communications Applications Conference; Antalya, Turkey; 2006. pp. 1-4.

[81] Arısoy E, Saraçlar M. Language modelling approaches for Turkish large vocabulary continuous speech recognition based on lattice rescoring. In: IEEE Signal Processing and Communications Applications Conference; Antalya, Turkey; 2006. pp. 1-4.

[82] Aksungurlu T, Parlak S, Sak H, Saraçlar M. Comparison of language modelling approaches for Turkish broadcast news. In: IEEE Signal Processing and Communications Applications Conference; Zonguldak, Turkey; 2016. pp. 1-4.

[83] Çiloğlu T, Çömez M, Şahin S. Language modelling for Turkish as a agglutinative languages. In: IEEE Signal Processing and Communications Applications Conference (SIU); Aydın, Turkey; 2004. pp. 1-2.

[84] Arısoy E, Can D, Parlak S, Sak H, Saraçlar M. Turkish broadcast news transcription and retrieval. IEEE Transactions on Audio, Speech, and Language Processing 2009; 17 (5): 874-883. doi: 10.1109/TASL.2008.2012313

[85] Hori T, Cho J, Watanabe S. End-to-end speech recognition with word-based RNN language models. arXiv 2018; arXiv:1808.02608 [cs.CL]. doi: 10.1109/SLT.2018.8639693

[86] Graves A, Jaitly N. Towards end to end speech recognition with recurrent neural networks. In: International Conference on Machine Learning; Beijing, China; 2014. pp. 1764-1772.

[87] Çelıktaş H, Hanılçı C. A study on Turkish text-dependent speaker recognition. In: IEEE Signal Processing and Communications Applications Conference; Antalya, Turkey; 2017. pp. 1-4.

[88] Yılmaz O, Saraçlar M. Audio diarization for Turkish broadcast news transcription. In: IEEE Signal Processing and Communications Applications Conference; Antalya, Turkey; 2011. pp. 379-382.

[89] Atalay NB, Oflazer K, Say B. The annotation process in the Turkish treebank. In: Proceedings of the 4th Linguistically Interpreted Corpora LINC-03; Budapest, Hungary; 2003. pp. 1-6.

[90] Eryiğit G, İlbay T, Arkan-Can, O. Multiword expressions in statistical dependency parsing. In: Proceedings of the second Workshop on Statistical Parsing of Morphologically-Rich Languages; Dublin, Ireland; 2011. pp. 45-55.

[91] Oflazer K, Say B, Hakkani-Tür DZ, Tür G. Building a Turkish Treebank, Invited Chapter in Treebanks: Building and Using Parsed Corpora. Dordrecht, Netherlands: Kluwer Academic Publishers, 2003.

[92] Strassel S, Walker K, Jones K, Graff D, Cieri C. New resources for recognition of confusable linguistic varieties: the LRE11 corpus. In: The Speaker and Language Recognition Workshop; Odyssey, Singapore; 2012. pp. 202-208.

[93] Craig G, Martin A, Graff D, Walker K, Jones K et al. The 2011 NIST Language Recognition Evaluation Plan (LRE11). Philadelphia, PA, USA: LDC, 2018.

[94] Arısoy E, Can D, Parlak S, Sak H, Saraçlar M. Turkish broadcast news speech and transcripts transcription and retrieval. IEEE Transactions on Audio, Speech and Language Processing 2009; 17 (5): 874-883. doi: 10.1109/TASL.2008.2012313

[95] Oflazoğlu Ç, Yıldırım S. Turkish emotional speech database. In: IEEE Signal Processing and Communications Applications Conference; Antalya, Turkey; 2011. pp. 1153-1156.

[96] Salor Ö, Pellom BL, Çiloğlu T, Demirekler M. Turkish speech corpora and recognition tools developed by porting SONIC: towards multilingual speech recognition. Computer Speech and Language 2007; 21 (4): 580-593. doi: 10.1016/j.csl.2007.01.001