symmetry

MDPI

Article

Development of Speech Recognition Systems in Emergency Call Centers

Alakbar Valizada [1,2,*], Natavan Akhundova [1,3] and Samir Rustamov [3,4]

[1] Artificial Intelligence Laboratory, ATL Tech, Jalil Mammadguluzadeh 102A, Baku 1022, Azerbaijan; natavan.akhundova@atltech.az
[2] Information and Telecommunication Technologies, Azerbaijan Technical University, Hussein Javid Ave. 25, Baku 1073, Azerbaijan
[3] School of Information Technologies and Engineering, ADA University, Ahmadbey Aghaoglu Str. 11, Baku 1008, Azerbaijan; srustamov@ada.edu.az
[4] Institute of Control Systems, Bakhtiyar Vahabzadeh Str. 9, Baku 1141, Azerbaijan
* Correspondence: alakbar.valizada@atltech.az

Abstract: In this paper, various methodologies of acoustic and language models, as well as labeling methods for automatic speech recognition for spoken dialogues in emergency call centers were investigated and comparatively analyzed. Because of the fact that dialogue speech in call centers has specific context and noisy, emotional environments, available speech recognition systems show poor performance. Therefore, in order to accurately recognize dialogue speeches, the main modules of speech recognition systems—language models and acoustic training methodologies—as well as symmetric data labeling approaches have been investigated and analyzed. To find an effective acoustic model for dialogue data, different types of Gaussian Mixture Model/Hidden Markov Model (GMM/HMM) and Deep Neural Network/Hidden Markov Model (DNN/HMM) methodologies were trained and compared. Additionally, effective language models for dialogue systems were defined based on extrinsic and intrinsic methods. Lastly, our suggested data labeling approaches with spelling correction are compared with common labeling methods resulting in outperforming the other methods with a notable percentage. Based on the results of the experiments, we determined that DNN/HMM for an acoustic model, trigram with Kneser–Ney discounting for a language model and using spelling correction before training data for a labeling method are effective configurations for dialogue speech recognition in emergency call centers. It should be noted that this research was conducted with two different types of datasets collected from emergency calls: the Dialogue dataset (27 h), which encapsulates call agents' speech, and the Summary dataset (53 h), which contains voiced summaries of those dialogues describing emergency cases. Even though the speech taken from the emergency call center is in the Azerbaijani language, which belongs to the Turkic group of languages, our approaches are not tightly connected to specific language features. Hence, it is anticipated that suggested approaches can be applied to the other languages of the same group.

Keywords: speech recognition; GMM; HMM; DNN; Kaldi; call center

Citation: Valizada, A.; Akhundova, N.; Rustamov, S. Development of Speech Recognition System in Emergency Call Centers. *Symmetry* **2021**, *13*, 634. https://doi.org/10.3390/sym13040634

Academic Editor: Tomohiro Inagaki

Received: 14 March 2021
Accepted: 4 April 2021
Published: 9 April 2021

1. Introduction

The job of call center agents can be overwhelming considering that they have to both talk on the phone and log the case information at the same time. There are many applications specifically designed to ease the job for call center agents. They can be connected via internet telephony, can include interactive voice response and other technologies to ensure a smooth experience.

New opportunities arose when Artificial Intelligence (AI) started to be applied to a lot of spheres of industry, as well as call centers. With cutting-edge AI technologies like speech recognition, call centers can now benefit from the automation of processes. Speech recognition is a capacity of a system to turn audio data of natural speech into a

corresponding text. A system with such a feature can transcribe the conversation, take insights from the call and perform actions while an agent is talking with a customer. Speech recognition seems to be a fast, easy and hands-free experience for agents. It is all due to the fact that people can speak almost four times more words in a minute than typing.

Additionally, the call system with speech recognition is capable of determining different accents, age and emotions of customers. Since a procedure of recording information of a customer and a case can be automated, agents will save more time and energy for other important organizational tasks. For example, customers of emergency systems are usually emotional and stressed. With speech recognition running in the background, agents can focus more on the customers and calm them down.

The task of recognizing speech in call centers is not an easy one. Call center data can be noisy, have a large vocabulary and various accents. It is also considered fluent and conversational speech. Therefore, considering all these problems, thorough investigation should be conducted to find the beneficial parameters for the collection of data, format and training methodologies.

Our study involved an investigation of key factors to focus on when developing a speech recognition system for a call center. The experiments were conducted and closely observed in order to explore a suitable approach for the training of the system. First, we introduce our data, which were collected by the Emergency Call Center of Azerbaijan. One dataset was collected from real calls and consists of dialogue speech, where order of the words in sentences is not correctly preserved, the language model is specific and the environment is noisy most of the time. Another dataset consists of summaries of these calls and is a monologue speech, where sentences are complete in a grammatically correct order and the environment is not noisy. Moreover, word count and vocabulary size are depicted per hour. Then, we describe the conducted experiments in detail. The purpose of the experiments consists of identifying training methods, text formatting ways and language models that will be the most favorable in the given context. All of the training procedures were conducted on Kaldi ASR, an open-source toolkit intended for speech recognition. We conducted training with Gaussian Mixture Model/Hidden Markov Model (GMM/HMM) and Deep Neural Network/Hidden Markov Model (DNN/HMM) methodologies on two different datasets. Afterwards, results are shown and discussed. As vocabulary size changes, outcomes of the tests change, as well, leading to use of different parameters to provide a better result.

Furthermore, speech recognition projects specified at call centers in Azerbaijani do not exist. Speech recognition systems available to us in Azerbaijani are only a local state project Dilmanc and Google Speech-to-Text. None of them show sufficient performance of speech recognition in the call centers because their train datasets were not built on dialogue data. Their language models do not justify themselves in the dialogue system, and there are no accents and also emotions in the speech. In fact, when tested with Google Speech-to-Text, the word error rate was 42.14% for the Summary dataset and 76.48% for the Dialogue dataset. This can be due to data having a specific language model, noisy environment and, overall, the language being a low-resource one, hence, the available platforms performed worse.

Even though the data are in the Azerbaijani language, our approaches are not tightly connected to specific language features. Therefore, the suggested methods can be also applied to other languages belonging to the Turkic group.

2. Literature Review

Based on the literature review, it can be concluded that speech recognition in call centers has not been widely investigated. The main reason could be that dialogue data of call centers, which can belong to private companies, may not be accessible for researchers. For the case of this paper, training Automatic Speech Recognition (ASR) based on common approaches from the studies with non-dialogue datasets does not give sufficient results.

The project Decoda is concentrated on reducing the cost of speech analytics systems development [1]. By proposing speech mining tools and methods, the authors aimed to reduce manual annotation process. They have described the corpus collected by the call-center of the Paris public transport authority. In this research, speech transcripts, semantic and syntactic annotation of dialogues have been analyzed; however, speech recognition was not performed.

In [2], Dr. Espana-Bonet and Dr. Fonollosa compared GMM/HMM with DNN/HMM for impaired speech using different architectures. The conclusion was that the neural network outperformed the Gaussian mixture model. We have achieved similar results, however, it should be noted that our data structure is not similar to the one in the study, nor are the data compatible with the language model in the study.

Authors of [3] conducted a research to change traditional a universal background model based i-vectors to HMM based i-vectors. The system uses HMM state alignment information to estimate i-vectors. By applying different techniques on experiments, the authors achieve 5–7% improvement on the word error rate.

Authors of [4] experimented with the method of lattice-free maximum mutual information adding end-to-end training with neural networks at one stage. The training was performed on a large vocabulary and gave better results than other end-to-end methods. The authors used Switchboard, a telephone speech corpus consisting of 300 h of speech in English, in their research. Their experiments show that the Word Error Rate (WER) in dialogue speech is higher than monologue speech.

In [5], the authors have suggested three probabilistic ways of creating transcripts for languages that have scarce data—either written transcripts or speech. They have concluded that the self-trained ASR outperformed cross-lingual baseline method, and the mismatched crowdsourcing method scored better than the self-trained one. This method is effective when transcripts of a low-resourced language are not available. Furthermore, the method results in lower accuracy compared to a human annotated dataset.

A new low-cost method for data augmentation was presented at [6]. The researchers of the study propose to alter the speed of the audio data by slowing down and speeding up. Experiments on four different tasks with data ranging from 100 to 960 h; an average improvement of 4.3% was achieved. The authors have used different speech corpuses, and one of them, being similar to ours, is a dataset of telephone conversations with a WER around 17%.

Recurrent neural networks are the main focus of [7] where authors investigated the RNN encoder–decoder approach. This approach encodes acoustic signals to feature vectors and decodes them to words. The approach was used to train 300 h of data and achieved promising results without any explicit language model. However this method is reliable for large amounts of data and shows relatively lower accuracy in small datasets.

Authors of [8] applied deep neural network to recognition of Kazakh speech using Kaldi speech recognition tool. The language, which is similar to Azerbaijani, was trained with 76 h of data. Specific features of the language were utilized to achieve an optimal result.

Design of speech recognition tool—Kaldi is described in [9]. Authors conducted several experiments with well-known databases to compare two speech recognition tools—HTK and Kaldi, as well as different training methods. They have also described the structure of the toolkits.

A system for analysis and monitoring performance at call centers is presented in [10]. Using speech recognition, the system transcribes the call, and with mining of the transcription, it acquires insights for agents and administrators. The researchers have used WebSphere Voice Server as a speech recognition tool, and its accuracy is less than 70%, which is not at the desired level.

Utilizing existent spoken language corpora for a specific task is a main focus of the study in [11]. Combining it with a task-specific data consisting of key words and phrases decreased the error rate of 13%, even though new data is existent and had different

stylistic differences. Central point of the research is to eliminate language model differences between different types of call center dialogue speeches. The acoustic difference was not considered in this study.

The authors of [12] have conducted research on improving ASR on 51 low-resource languages with a multilingual model. They have reached more than 20% reduction in average WER using three different methods, particularly joint model, joint model with a language input and multi-head model. This research is focused on the acoustic model and no language model or data labeling methods were investigated.

In [13], authors experimented with a deep neural network via the adding data augmentation and ensemble method. The author's hypothesis was that these two methods have proven their effectiveness in machine learning and can positively contribute to the training process. The assumption proved to be right when experiments showed an increase in the performance of the system.

Comparison of two well-known automatic speech recognition tools—Kaldi and CMUSphinx were compared with a small dataset on three different criteria [14]. Kaldi has shown better results on accuracy and variance, whereas CMUSphinx completed training in less time.

Authors from Amity University studied approaches and techniques for improvement of ASR performance in call centers [15]. They suggested an approach of using emotion recognition to measure customer satisfaction with speech recognition. The authors have experimented with tools like Sphinx, Google and Microsoft SDK for speech recognition, however these systems show low accuracy for our data.

Dr. Seltzer, Dr. Yu and Dr. Wang conducted a study to test neural network systems in noisy environments in [16]. Their three approaches were: (1) training with multi-conditional data; (2) using feature enhancement to remove distortions in the observations prior to training; (3) incorporating a noise model or noise estimate into the network itself. They have concluded that a DNN-based acoustic model performs much better than the GMM-HMM state-of-the-art algorithms and the robustness of DNN can be improved with noise-aware training by 7.5%.

The authors of [17] conducted a study of training an acoustic model where acoustic units would be whole words. They used a model of LSTM-based RNNs with CTC loss. Semi-supervised training was carried out on 125,000 h of data. The authors claim that their word-based system performs better than a phone-based system.

In [18], authors from Cambridge University claimed that the use of CMLLR-based speaker adaptive training for a jointly MPE trained Tandem system is more accurate than the conventional Tandem systems. They have achieved 4% of decrease in WER results on the jointly trained Tandem SAT system.

In [19], the authors investigated pattern matching techniques for speech recognition in noisy and noiseless environments. They also discussed feature extraction methods. The article is a review of different studies on the topic.

Dr. Zhang, conducted a study to train and compare hidden Markov model with Gaussian mixture model, deep neural network and deep belief network with the well-known TIMIT database [20]. According to word error rate results, DBN performed better than the other two speech recognition systems. Our experiment was conducted using HMM/GMM and HMM/DNN models, nevertheless, our dataset encapsulates dialogue data, which have a different structure than the TIMIT dataset.

Each of these methods—DNN, DBF, LSTM and HMM—outperform each other under specific conditions and train sets. However, these methods were not applied to recognize dialogues in the Azerbaijani language. As the datasets are specific and in low-resource languages, we have not observed different methodologies being applied, such as the various labeling methods. In this study, 27 h of call and 53 h of summary speech data from a call center were labeled and trained by us based on different types of GMM/HMM and DNN/HMM models. To increase the accuracy of speech recognition systems, various data labeling and recognition methods were researched.

3. Data Preparation

Two distinct datasets were collected from the Emergency Call Center. The first one is the main dataset consisting of dialogue speech and the second dataset consists of voiced summaries of calls. As the study is focused on developing and improving ASR for dialogue speech, most of the experiments were conducted on the main data, and the alternative one was used to compare and observe differences. To give a more detailed overview of the main data: it has 4333 different dialogues with 40 female call center agents' speech in the Azerbaijani language. In total, the data are 27 h of call recordings and are about emergency cases like fire, locked doors, gas, drowning, etc. The purpose of extracting only agents' speeches from audio recordings is to achieve precise recognition for summaries of dialogues and information about cases. The calls were recorded in such a way that the left channel of audio recordings is the agent's voice and the right channel is the client's voice. Recordings were cut into small audio tracks to get better phoneme alignment and transcribed. Sound files are stereo WAVs with 8000 sample rate, 16-bit precision, 128 k bit rate. For splitting, we had used sox audio utility [21].

The first step of splitting is trimming silence from the beginning of the audio. There are two significant parameters for silence: duration and threshold. Duration indicates the amount of time that non-silence must be detected before it stops trimming audio. The threshold is used to indicate what sample value you should treat as silence and is used for background noise. For this case, the duration was 0.4 s and the threshold was 0.05% of volume.

The second step after trimming silence from the beginning of the audio is to remove all audio after silence is detected. For this process, we took the silence duration as 0.2 s and the threshold as 0.05% of volume.

The third step is about repeating the first and second steps recursively until the end of audio.

Data Analysis

During the collection of the recordings, it can be observed that a vocabulary steadily increased at each hour-frame (Figure 1); however, starting from 15 h of data, the number of newly added words per hour decreased from 700 to 400 words. It is predicted that the number will reach stability at some point and after that, adding new transcriptions to the dataset will have less value in terms of vocabulary.

Additionally, 61% of vocabulary was used only once due to the fact that the contents of calls are diverse. It indicates that entities from calls, including names, surnames, street addresses, etc, are various, and as a result, it affects the vocabulary. These words correspond to only 5% of all words.

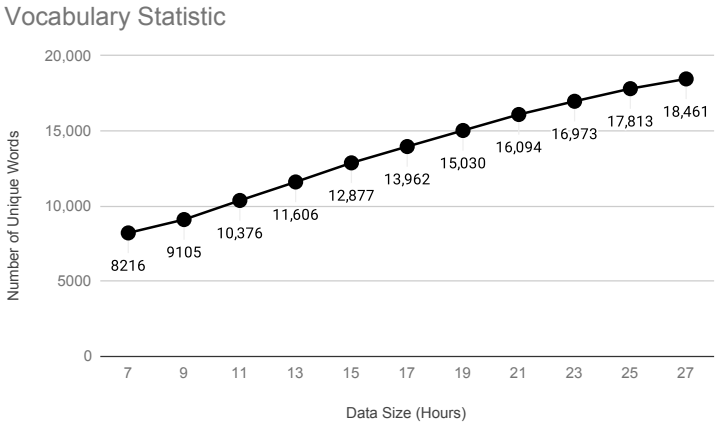

Figure 1. Vocabulary statistic through hours.

The total number of words were increasing but with relatively higher speed. The rate of aggregating new words was between 10,000 and 20,000 words per hour (Figure 2).

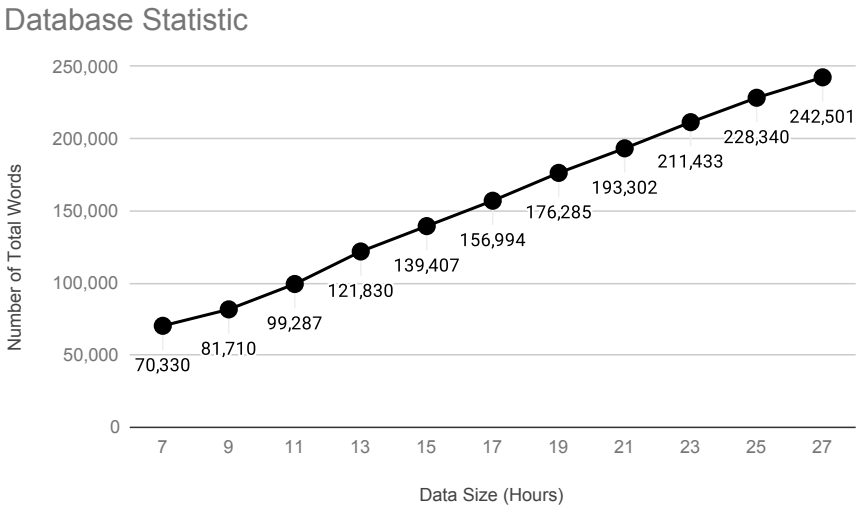

Figure 2. Database statistics through hours.

According to Zipf's law, the plotting frequency of words in a large corpus against their rank in a log scale gives a straight line. It is worth knowing whether this corpus of dialogue system satisfied the law. Similar to the description in [22], Figure 3 depicts that the first part of the curve is unsmooth, the middle part is smoother and the last part—from 1000 to 10,000—is angled downward.

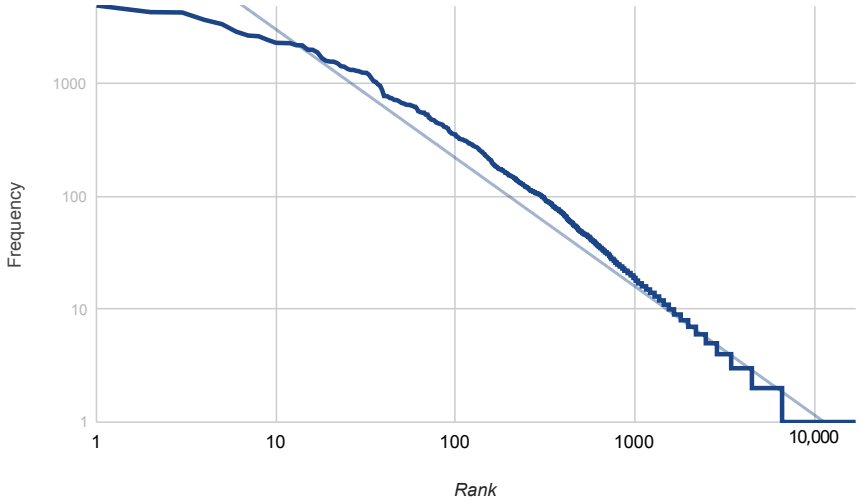

Figure 3. The frequency–rank scale of words in the database of 27 h.

Additionally, it is important to note that each speaker has contributed in different ratios. Total time of speech for one speaker changes from 7 to 91 min. Figure 4 depicts these differences for each speaker ID.

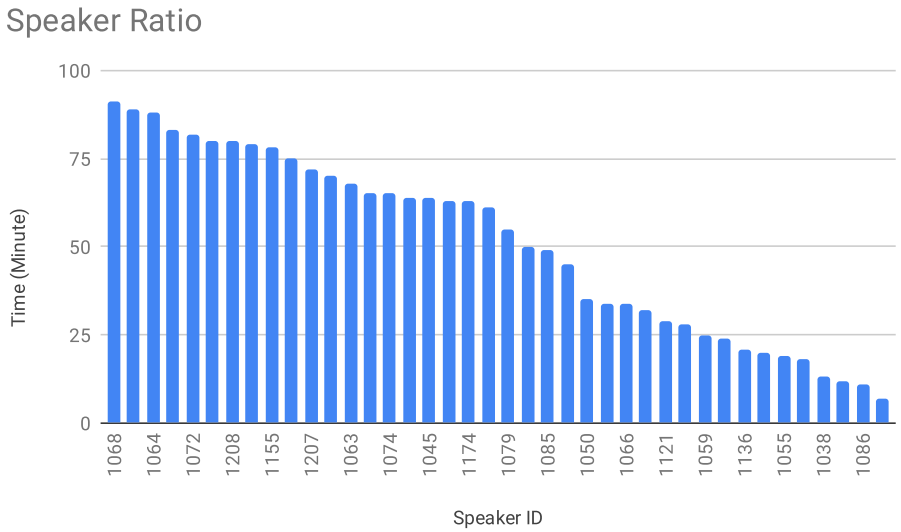

Figure 4. Amount of minutes each speaker has spoken in the dataset.

Another dataset was collected to compare methodologies between two different datasets. This dataset consists of 53 h of summaries of emergency dialogues voiced by call center agents. It has 39,118 transcripts, 396,933 total words and 16,546 unique words, which is 1.5 times more than the previous dataset. The transcripts differ from the main dataset by their grammatically correct order of words. Compared to the main one, this dataset contains complete sentences and less grammatical errors in words.

4. Methodology

4.1. Feature Extraction

The pre-step of building and training models is extracting features from audio signals, which are Mel Frequency Cepstral Coefficients (MFCC). The purpose of MFCC extraction is to derive significant information from audio signals. For each defined frame, the windowing function is applied and energies are calculated [23]. The procedure continues with taking the Fourier transform of a signal. Then obtained results are passed through the Mel filter, to which inverse Fourier transformation is applied.

After extracting MFCC from the audio data, the monophone model was trained. The model acts as a building block for other more complex models and does not include any information about preceding or following phones [24]. In Kaldi, in order to adjust the parameters of an acoustic model, an alignment procedure was performed. This step aligns audio to the transcripts and allows the next model to start improving the output of the process. Each of the models that are compared in this study depends on the previous model's alignment. Thus, monophone training and its alignment are essential for conducting the first training method.

4.2. Acoustic Models

It is widely known in the speech recognition field that two significant factors influence recognition performance, which are acoustic and language models. In this section, the aim is to define the most accurate acoustic model for dialogue systems.

The monophone model is not effective due to low accuracy rate. Therefore, training methods are chosen with higher phoneme states for the current problem. Five different training methods of HMM and a method of DNN together with HMM were used for experimenting and improving accuracy rates. The first method is a simple triphone training named TRI1. This training takes one step over monophone training considering both preceding and following phonemes.

The second method—TRI2B—uses Linear Discriminant Analysis (LDA) with Maximum Likelihood Linear Transform (MLLT). LDA generates HMM states from feature

vectors with reduced feature space. This space is used by MLLT, which normalizes speakers with a sole transformation [24].

TRI2B + MMI is the next method, which includes in itself TRI2B with Maximum Mutual Information (MMI). MMI, being a sequence discriminative training criteria, is the information that belongs to both word sequences and distributions of the observation [25].

The other method is LDA together with MLLT and Speaker Adaptive Training (SAT), which is named TRI3B. SAT is not completely independent and not completely dependent on speaker information. It is an approach where speaker-independent training is adapted to a speaker using its data [26]. With data transformation, it normalizes speaker and noise. Last but not least, the training method of HMM is TRI3B with MMI (TRI3B + MMI). It combines the TRI3B method with a sequence discriminative training criteria.

A method, slightly different from the others, is NNET3, which is based on Deep Neural Networks (DNN) together with HMM. The principle behind this methodology is to use many layers of non-linear hidden units for training and doing a forward–backward on a decoding graph which acts as an MMI model without lattices [27]. The objective function for a model is the log–probability of the correct phone sequence [11].

These six methodologies were tested in terms of recognizing unseen data where 90% of data were dedicated for training and the rest 10% for testing. Besides, testing of unseen speakers is also conducted. It is performed as taking 90% of all speakers for training and 10% for testing. The results of both experiments are described using two language models—unigram and trigram. Unigram will help to clearly compare acoustic methodologies as it is a depiction of accuracy for pure acoustic signals. Trigram, on the other hand, is a depiction of accuracy when preceding and following words are taken into account and will help to compare methodologies when the language model is present in the system.

4.3. Language Models

Additionally, the methods were tested with different n-grams, starting from unigram and ending in five-grams. Assessment of the language models was carried out using intrinsic and extrinsic methods of quality evaluation.

A popular method of intrinsic quality evaluation is perplexity [28]. In this method of quality evaluation, a model can be measured with a metric without considering any particular context. Perplexity is calculated as the inverse probability of test set, normalized by number of words. Modified Kneser–Ney discounting and Witten–Bell discounting were utilized as smoothing methods. The smoothing methods were referenced from SRILM library, which is used in Kaldi.

Besides intrinsic methods for testing performance of language model, results were calculated using Word Error Rate (WER) and Sentence Error Rate (SER) metrics. WER is a percentage of wrongly inserted, deleted and substituted words by the tool during recognition, and SER is a percentage of wrongly recognized sentences by the tool [28].

4.4. Data Labeling Methods

The last part of the experiments focused on data labeling and formatting. The common approach for data labeling in speech recognition is to write transcriptions of audio signals in correct spelling. However, there could be a risk of bias emerging in forced alignments in small datasets. When pronunciation differs from the spelling of words, phonemes could be aligned to different sounds and recognized mistakenly in a testing process. For a large amount of data, this risk can be negligible, however, in this case, it is worth testing and comparing error rates.

In order to decrease error rates, four approaches of testing the dataset were conducted in the study. The question for the approaches was whether to perform spelling correction or not and if performed, whether to change words before training or after, or add different pronunciations of words in the dictionary. To answer these questions, 90% of all data were chosen to train the system, while the rest 10% were designated for testing.

For the first approach, there are no changes in the labeling of data, and the error rates are taken directly from the results of TRI3 + MMI and NNET3 training method. The name for this approach would be Raw (Train + Test).

The other method is an advised one described in the documentation of Kaldi, where different pronunciations for a word are added to the dictionary. In this way, the tool takes into consideration different forms of the same word and calculates their probabilities. This testing method is named Standard.

The idea behind the next method is to correct misspelling words that differ in pronunciation in the training dataset before a training process. Some words in the database have many counterparts, such as "OPERATOR", "APERATIR", "OPERATIR". Others are cut in-between the words or silent at the beginning. This is due to the fact that call center agents transcribed audio tracks the way they heard it. These words are all corrected before training and testing. This approach is named SC (Train + Test). Naturally, the number of unique words decreased in this approach.

The problem about the previous method is that while correcting words, phonemes that were not spelled in the transcript are added. For example, "RATOR" in the transcript becomes "OPERATOR", artificially adding the "OPE" part, which did not exist in the track. There is a risk that it can influence the training process, misleading it.

The fourth method corrects spelling after the training process is finished. It is not mandatory to change the whole dataset to get results for this approach. The test dataset and recognition results of the tool are the main collection methods of data to perform spelling correction. The approach is named Raw (Train) + SC (Test).

5. Experimental Results

In this section, we put experimented with various methodologies for acoustic models, language models and data labeling methods. Firstly, to investigate acoustic models, we compared different HMM models with each other based on WER and SER metrics. After this step, we compared the best HMM acoustic model with the DNN model with the same metrics. Obtained results show that DNN outperforms all other HMM models. For the language model, we took different n-grams and evaluated them both in intrinsic and extrinsic ways. After conducting all experiments, we come to the conclusion that the trigram method for a language model is more robust and less error-prone to variations than other methods. Lastly, in the section about data labeling methods, we consider four data labeling methods and show their comparisons. Observing the comparison results, it can be argued that the "Raw (Train) + SC (Test)" is more suitable for call center applications.

5.1. Acoustic Models

To begin with, five training methods of HMM were tested through 7–27 h to find the best performing acoustic model for the current task. In order to clearly interpret the output, in terms of the acoustic model, a unigram language model was used.

As it can be seen in Figure 5, error results for the unigram was mostly more than 40%. The rates are slowly decreasing towards the end. According to the graph, TRI3B + MMI is the most accurate method at each milestone.

To see changes in accuracy results, the language model was added into the experiments. A closer look at the trigram statistics in Figure 5 can show that the order of accuracy of training methods in trigram is preserved as in the unigram model.

After comparing HMM training methods to find the most accurate and stable one, a DNN model was also trained for comparison. The method named as NNET3 was trained for 27 h with both unigram and trigram language models and compared with a HMM model in Table 1.

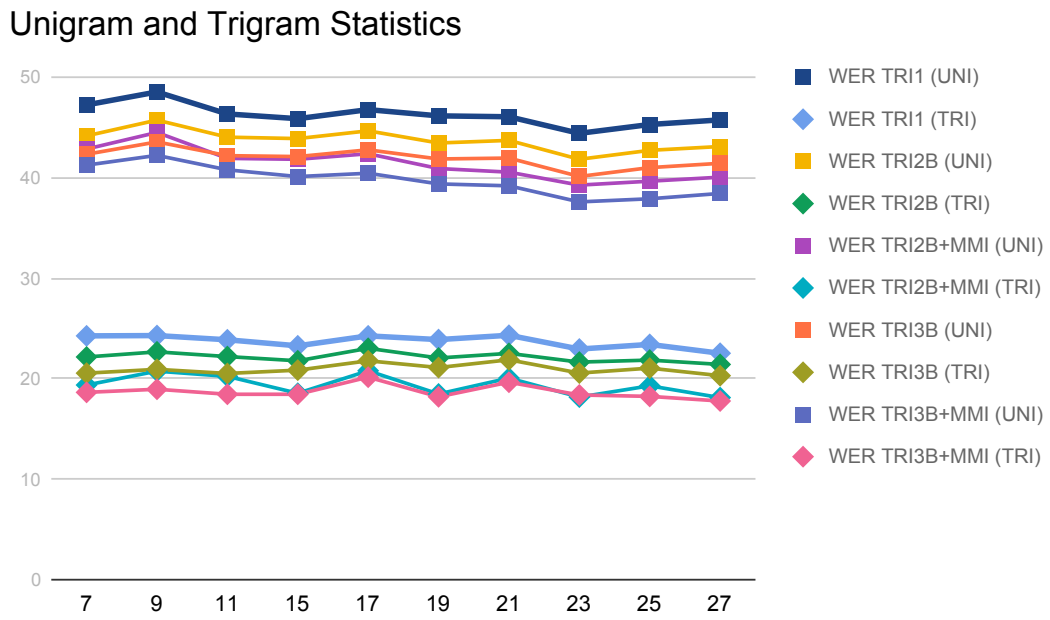

Figure 5. Unigram and Trigram statistics of Word Error Rate (WER) of the training methods through hours.

Table 1. Word error rates of the TRI3B + Maximum Mutual Information (MMI) method of the Hidden Markov Model (HMM) and the NNET3 method of HMM/Deep Neural Network (DNN) for 27 h.

Training Methods	Unigram	Trigram
TRI3B + MMI	38.42	17.77
NNET3	33.28	14.01

To see the mean and variance values of training methods, a 10-fold cross-validation of the trigram at 27 h of data was conducted, where ten different test sets were chosen randomly. Tables 2 and 3 show the values for the word error rate and sentence error rate of each training method, respectively.

Table 2. Mean and variance values of trigram word error rates of 10-fold cross-validation for 27 h.

Training Methods	Mean (WER)	Variance (WER)
TRI1	23.327	0.2022
TRI2B	21.905	0.1510
TRI2B + MMI	18.609	0.3483
TRI3B	20.936	0.2391
TRI3B + MMI	18.535	0.2308
NNET3	**14.148**	0.0367

Table 3. Mean and variance values of trigram sentence error rates of 10-fold cross-validation for 27 h.

Training Methods	Mean (SER)	Variance (SER)
TRI1	53.51	0.4546
TRI2B	51.845	0.2940
TRI2B + MMI	48.374	1.1952
TRI3B	50.647	0.7173
TRI3B + MMI	47.885	1.0422
NNET3	**42.156**	0.3588

The alternative dataset was also tested by each training methodologies with unigram and trigram as language models. The differences between two datasets regarding perfor-

mance of acoustic models can be seen in Figure 6. It can be observed by looking at the unigram results that error rates of call data start near 45%, while the ones of summary data start near 25%. Adding a language model improved the accuracy, but still kept the difference between these two datasets. Therefore, unigram values and trigram values for datasets differ by around 20% and 10%, respectively.

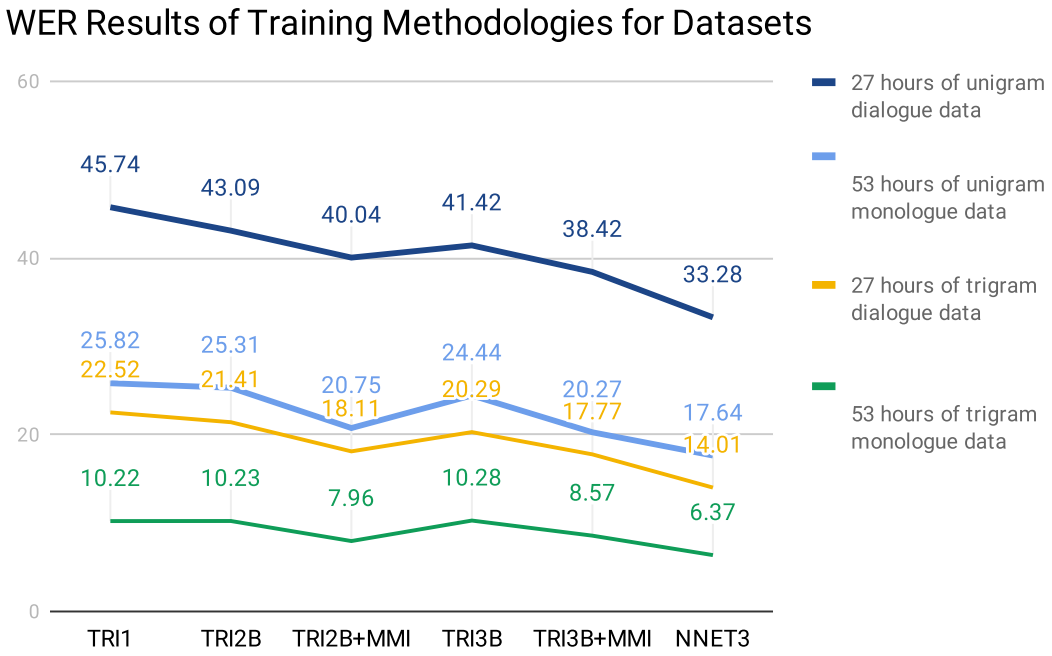

Figure 6. Unigram and Trigram statistics of WER by the training methods for 27 h of call data and 53 h of summary data.

5.2. Language Models

The second task is to find which n-gram is effective for dialogue systems. Since dialogue systems differ from other models, existing language models based on Wikipedia or news data do not justify themselves. A language model should be specific and relevant to the task of call center dialogue system.

First of all, intrinsic quality evaluation of the language model has been conducted. The results of perplexity are shown in Table 4, with the comparison of word error rate results of TRI3B + MMI and NNET3 models. Both perplexity values decreased until four-gram. At five-gram, the values were a little bit higher. The WER results decreased until that threshold and increased at five-gram.

Table 4. Comparative review of intrinsic and extrinsic evaluation methods.

Evaluation	Intrinsic		Extrinsic	
n-Grams /Methods	**Modified Kneser-Ney Discounting**	**Witten-Bell Discounting**	**WER TRI3B + MMI**	**WER NNET3**
1	768.4366	485.8573	38.42	33.28
2	124.8557	96.66415	18.81	14.37
3	96.03731	86.28279	17.77	14.01
4	79.63794	84.984	18.1	13.57
5	79.6429	84.98929	18.36	14.32

Additionally, to have a visual representation of differences of n-grams, WER results of all training methods were described for 27 and 53 h of data on Figures 7 and 8, respectively.

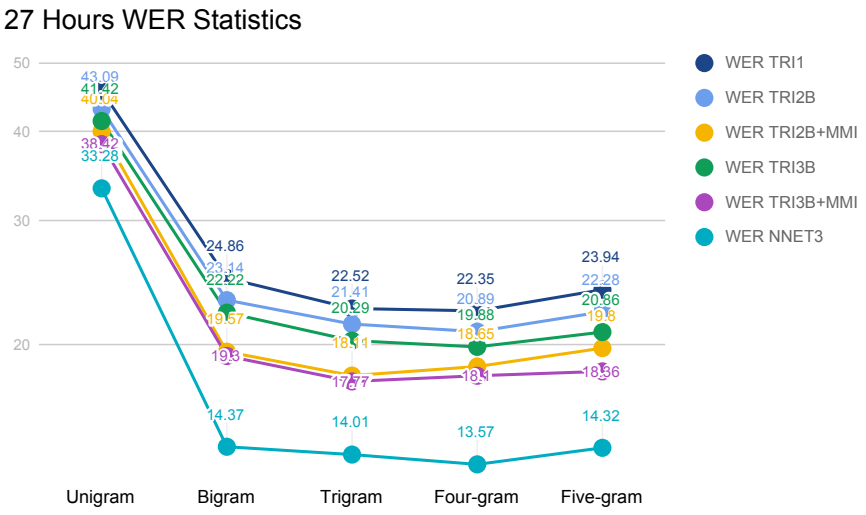

Figure 7. WER statistics of 27 h of data for six different training methodologies.

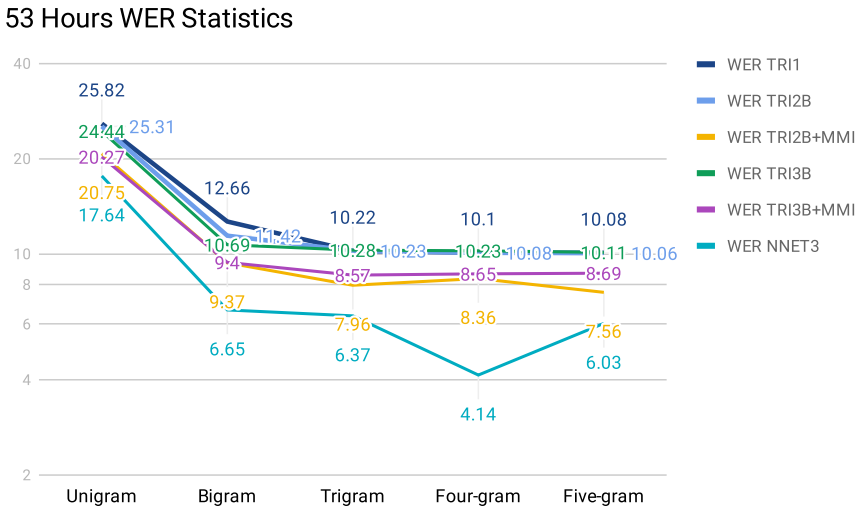

Figure 8. WER statistics of 53 h of data for six different training methodologies.

Starting from unigram, five n-grams were utilized to train and test the dataset. As it is observed in the figures, unigram shows the highest error rate for all training methods. Trigram, on the other hand, is a turning point of a decrease for all hours of data. After trigram, error percentages of methods are slightly decreasing at four-gram and slightly increasing at five-gram.

5.3. Data Labeling Methods

The third experiment was to compare the four data labeling approaches for the dataset. A training method with MMI and a method of DNN/HMM were chosen as they scored with the least error percentages in previous tests.

For the first approach, there were no changes in the labeling of data, and the error rates were taken directly from the results of training method. The name for this approach is Raw (Train + Test). The first test has the same error rate of 27 h of data when trained 10-fold cross-validation with three grams. It has a vocabulary of 18,461 unique words and mean WER of around 16% and mean SER of around 44% for both training methods.

The second method consists of adding and describing various pronunciations of a word to the lexicon. This is a common approach used in speech recognition. After the

change, the total number of words increased from 18,461 to 28,605, which means more than 10,000 pronunciation data have been given to the system to learn. The results have increased a little bit more than 1% for the TRI3B + MMI model and decreased around 1% for the NNET3 model, on average.

The next test is where spelling correction was applied to the dataset before training. The initial dataset contains 27 h of data, and its vocabulary consists of 16,221 words after spelling correction. It has decreased due to the fact that duplicates of the same word with different forms are erased in the vocabulary. After the cross-validation, the final result is that error rates for words of both training methods decreased to 13% on average. Sentence error rates have also decreased resulting values near 36%. The "Raw (Train) + SC (Test)" approach has the least word and error rates for both methods. In all of the cases with spell correction, NNET3 has less error values and also less variance. The results are depicted in Table 5.

Furthermore, these methods were tested with shuffling speaker voices, such as 10% of all speakers were treated as test data. Both of the methods have the lowest error rates for words and sentences in the fourth approach. The results can be observed in Table 6.

Table 5. Error rates of four data labeling tests conducted for shuffled 27 h data.

Name	**Raw (Train + Test)**		**Standard**		**SC (Train + Test)**		**Raw (Train) + SC (Test)**	
Methods	**TRI3B + MMI**	**NNET3**	**TRI3B + MMI**	**NNET3**	**TRI3B + MMI**	**NNET3**	**TRI3B +MMI**	**NNET3**
WER (%)	17.77	14.01	19.13	13.62	15.52	9.94	13.42	**9.08**
SER (%)	46.36	42.09	48.06	40.62	41.08	31.31	36.42	28.78

Table 6. Error rates of four data labeling tests conducted for 27 h of data with shuffled speakers.

Name	**Raw (Train + Test)**		**Standard**		**SC (Train + Test)**		**Raw (Train) + SC (Test)**	
Methods	**TRI3B + MMI**	**NNET3**	**TRI3B + MMI**	**NNET3**	**TRI3B + MMI**	**NNET3**	**TRI3B + MMI**	**NNET3**
WER (%)	20.13	14.18	18.37	14.02	15.85	10.11	15.26	**9.35**
SER (%)	50.98	41.81	47.84	41.33	42.34	31.82	40.54	29.48

The same approaches have been put under the test for the alternative dataset. Compared to the other results, the third method called SC (Train + Test) shows the highest accuracy where error rates are below 7% for WER and 36% for SER. Overall, the most accurate method was NNET3, as in other experiments. The results are expressed in Table 7.

Table 7. Error rates of four data labeling tests conducted for shuffled 53 h data.

Name	**Raw (Train + Test)**		**Standard**		**SC (Train + Test)**		**Raw (Train) + SC (Test)**	
Methods	**TRI3B + MMI**	**NNET3**	**TRI3B + MMI**	**NNET3**	**TRI3B + MMI**	**NNET3**	**TRI3B + MMI**	**NNET3**
WER (%)	8.57	6.37	8.34	6.17	6.89	5.37	8.52	6.01
SER (%)	41.84	33.34	41.03	31.81	35.29	29.02	41.61	32.02

6. Discussion

In this paper, the recognition of dialogue speech in call centers for emergency cases has been investigated. The main reasons for such datasets not being investigated widely and deeply is likely that call center data are sensitive, not publicly available and mostly belong to private and state companies. It is worth mentioning that the current work was carried out within the project of the Ministry of Emergency Situations of the Azerbaijan Republic. It is primarily dedicated to obtaining highly accurate and robust speech recognition via selecting effective techniques out of many acoustic, language and labeling methodologies.

6.1. Acoustic Models

The first research question in the current study is: which acoustic model is effective for dialogue data? To answer this question, six different GMM/HMM and DNN/HMM model configurations were tested. Similar techniques can also be observed by [2,3,8,13,16,20]; however, the datasets in these studies are telephone call data or common speech data. Nevertheless, the results obtained in the current study are very similar to those results.

In all of the predefined milestones, TRI3B + MMI is the most accurate method for HMM models. As a next step, adding DNN to models resulted in an above 4% decrease for WER and almost 6% increase for SER in accuracy. Comparing training methods between the least accurate and the most accurate one, the performance increased 9.2% for word accuracy and 11.4% for sentence accuracy, according to the mean values. In [7], authors investigated the RNN encoder–decoder approach and achieved promising results, however, this shows relatively lower accuracy in small datasets.

When it comes to variance, the least varied method is NNET3 and the most volatile is TRI2B+MMI. Overall, sentence accuracy is more volatile in comparison with word accuracy. The authors of [20] did similar experiments with the well-known TIMIT database. They came to the conclusion that the deep belief network performed better than the other two speech recognition systems. In contrast, our dataset encapsulates dialogue data that has a different structure than the TIMIT dataset.

When it comes to the alternative dataset, it can be concluded that the most accurate methodologies keep the order in the main dataset, however, in the case of the trigram model experiment, TRI1 is more accurate than TRI2B and TRI3B. Nevertheless, TRI3B + MMI was the most accurate for HMM and NNET3 was more accurate than HMM methods in all cases. In [2], the authors also compared GMM/HMM with DNN/HMM for impaired speech. We have achieved similar results, however, it should be noted that our data structure is not similar to the one in the study, nor are the data compatible with the language model in the study.

Taking into account the results and performances, NNET3 will be used for future studies.

6.2. Language Models

The second research question is which language model is effective for dialogue speeches. In order to address this question, n-gram models from 1 to 5 have been tested and evaluated with intrinsic and extrinsic methods.

In most of the cases, error rates are rapidly decreasing towards trigram. After trigram, rates are slightly less and can be considered as slightly stable (Figure 7). Error rates for words mostly decreased from over 40% to 20%, and error rates for sentences declined from over 70% to 50%.

On the alternative dataset—the Summary dataset—the higher the n-gram, the better is the result. This can be explained by the fact that the dataset only contains grammatically correct sentences, therefore, n-gram contributes to the accuracy more than to the main dataset. Compared to the main dataset, TRI2B and TRI2B + MMI scored better than its counterparts with speaker adaptive training; however, NNET3 has the least error rates.

To conclude, the accuracy order of training methods does not change a lot with the trigram model within different hours, and it also has less word and sentence error rates than other models. Furthermore, by analyzing similarities between two datasets regardless of their obvious differences, trigram is more desirable to work with.

6.3. Data Labeling Methods

Most of the research works [1,6,13] were devoted to automatize labeling methods rather than using effective labeling to get higher accuracy in the recognition process. In this study, spelling correction was applied before training procedure to increase performance of the system.

Applying spelling correction to transcripts before the training resulted in the word error rate decreasing 2% for TRI3B + MMI, and 4% for NNET3. It also has decreased

5% for TRI3B + MMI and 11% for NNET3, in terms of SER. In terms of the alternative data, the word accuracy has also increased 1%, so as the sentence accuracy, which scored 4–7% more. The SC (Train + Test) approach outperforms other labeling methods for the alternative data and also justifies itself as both numbers for WER and SER for each dataset were reduced. It also has removed duplicate words and reduced vocabulary size.

The outcomes of the fourth method—"Raw (Train) + SC (Test)"—shows the best results and outperforms the other three methods for the main dataset. In the approach, WER has reduced by 4–5% and SER decreased 10–13%, in comparison to the first method, however NNET3 has less error rate at the end. In addition, for the alternative dataset, the decrease in error rates was 1%. The reason why the fourth method performs better for the main dataset is that grammar and orthographic rules are applied after the training. In this way, the correction of words does not mislead the training of phonemes and also removes wrongly spelled counterparts of words, leaving a smaller vocabulary.

Proceeding with the experiments, in Table 6 where speaker accuracy has been tested for the main dataset, both of the measurements have decreased towards the fourth column. In the Standard method, error rates have decreased by only 1%. The SC (Train + Test) approach performs much better by decreasing WER 4% and SER 8–10%. The lowest score belongs to NNET3 training methodology.

In Raw (Train) + SC (Test), the decrease is 5% for WER and 10–12% for SER. NNET3 scored the lowest error rate in this test too. The approach has the highest decrease ratio among other approaches.

In summary, there are two apparent outcomes derived from these tests. The first outcome is that comparing two different performances of TRI3B + MMI, it performed better at recognizing new speakers due to speaker adaptive training. On the other hand, DNN/HMM performed the best for both tasks. Secondly, even though correct spelling of transcripts is good for large data, for the current issue of dialogue systems, it creates notable differences of accuracy when applied after the training procedure of the dataset.

7. Conclusions

This paper proposed an investigation of different models and methods for ASR in emergency call centers especially for improving the quality and performance of ASR for the Azerbaijani language. The noticeable approach which differs our work from others is comparison of data labeling methods. In this paper, we showed how the suggested data labeling approach with spelling correction result in outperforming other methods by a notable percentage.

It should be emphasized that none of the existing available Azerbaijani speech recognition systems show sufficient performance for call center data. In fact, when tested with Google Speech-to-Text, word error rate was 42.14% for Summary Dataset and 76.48% for Dialogue Dataset. This can be due to data having a specific language model, noisy environment and, overall, the language being low-resource one, hence, available platforms perform lower. Since recognition systems for Azerbaijani dialogue speeches are not available, this research was conducted to develop a speech recognition system which is built on dialogue data. The most accurate language model, training and data labeling methods were identified via HMM and DNN/HMM testings of Azerbaijani audio data from the emergency call center.

Based on the results, the most precise training for the acoustic model in terms of HMM can be reached with methods that utilize Maximum Mutual Information. Particularly, Speaker Adaptation Training with Maximum Mutual Information gave the least error rates than others. Nevertheless, the DNN/HMM method has performed the best in acoustic model experiments.

The most accurate and the least risky language model is trigram for both call and summary datasets. Through hours, accuracy rates of datasets show better results when reached trigram, and no overfitting risk can be observed at this n-gram.

It can be concluded that when the dataset is comparably small, formatting with correct spelling can be applied on datasets to increase accuracy. Based on the results, spell correction after the training process shows better performance for the call dataset where correct order of words in sentences is not preserved. It is when the system has learned phonemes without misleading, and further changes to the transcripts brings more accuracy. Training with a combination of linear discriminant analysis, maximum likelihood linear transform and maximum mutual information outperforms other HMM methods when datasets are correctly spelled. Furthermore, it can be inferred that speaker adaptation training with maximum mutual information can perform better when recognizing new speakers than other training methodologies, except the deep neural networks method, which achieved the highest accuracy in every experiment.

Future experiments could include the bidirectional language models and transformer-based acoustic models to find the parameters for the most accurate models.

Author Contributions: Conceptualization, S.R., A.V. and N.A.; methodology, A.V. and N.A.; validation, S.R.; formal analysis, S.R.; investigation, A.V. and N.A.; resources, A.V.; data curation, A.V. and N.A.; writing—original draft preparation, N.A.; writing—review and editing, S.R. and N.A.; visualization, A.N; supervision, S.R.; project administration, S.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by ATL Tech LLC.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Due to sensitive information about emergency cases in the country, research data will remain confidential.

Acknowledgments: This work was carried out in the Center for Data Analytics Research at ADA University and in Artificial Intelligence Laboratory at ATL Tech.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASR	Automatic Speech Recognition
HMM	Hidden Markov Model
GMM	Gaussian Mixture Model
DNN	Deep Neural Network
WER	Word Error Rate
SER	Sentence Error Rate

References

1. Bechet, F.; Maza, B.; Bigouroux, N.; Bazillon, T.; El-Beze, M.; De Mori, R.; Arbillot, E. DECODA: A Call-Center Human-Human Spoken Conversation Corpus. In Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey, 21–27 May 2012. Available online: http://www.lrec-conf.org/proceedings/lrec2012/pdf/684_Paper.pdf (accessed on 1 January 2021).
2. Espana-Bonet, C.; Fonollosa, J.A.R. Automatic Speech Recognition with Deep Neural Networks for Impaired Speech. In *Advances in Speech and Language Technologies for Iberian Languages*; Springer: Cham, Switzerland, 2016; pp. 1–11. [CrossRef]
3. Garimella, S.; Mal, A.; Strom, N.; Hoffmeister, B.; Matsoukas, S.; Parthasarathi, S.H.K. *Robust i-Vector Based Adaptation of DNN Acoustic Model for Speech Recognition*; Interspeech: Shanghai, China, September 2015. Available online: https://s3-us-west-2.amazonaws.com/amazon.jobs-public-documents/2015_ivector_paper.pdf (accessed on 1 January 2021).
4. Hadian, H.; Sameti, H.; Povey, D.; Khudanpur, S. *End-To-End Speech Recognition Using Lattice-Free MMI*; Interspeech: Shanghai, China, 2018. [CrossRef]
5. Hasegawa-Johnson, M.A.; Jyothi, P.; McCloy, D.; Mirbagheri, M.; Di Liberto, G.M.; Das, A.; Ekin, B.; Liu, C.; Manohar, V.; Tang, H.; et al. ASR for Under-Resourced Languages From Probabilistic Transcription. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 50–63. [CrossRef]

6. Ko, T.; Peddinti, V.; Povey, D.; Khudanpur, S. *Audio Augmentation for Speech Recognition*; Interspeech: Shanghai, China, September 2015. Available online: https://www.danielpovey.com/files/2015_interspeech_augmentation.pdf (accessed on 1 January 2021).
7. Lu, L.; Zhang, X.; Cho, K.; Renals, S. *A Study of the Recurrent Neural Network Encoder-Decoder for Large Vocabulary Speech Recognition*; Interspeech: Shanghai, China, September 2015. Available online: https://homepages.inf.ed.ac.uk/srenals/ll-rnn-is15.pdf (accessed on 1 January 2021).
8. Mamyrbayev, O. Automatic Recognition of Kazakh Speech Using Deep Neural Networks. In *Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2019; Volume 11432. [CrossRef]
9. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L. The Kaldi Speech Recognition Toolkit. 2011. Available online: https://www.researchgate.net/publication/228828379_The_Kaldi_speech_recognition_toolkit (accessed on 1 January 2021).
10. Mishne, G.; Carmel, D.; Hoory, R.; Roytman, A.; Soffer, A. Automatic Analysis of Call-Center Conversations. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management—CIKM 05, Bremen, Germany, 31 October–5 November 2005. [CrossRef]
11. Miki, K.; Hatazaki, K.; Hattori, H. Efficient Language Model Development for Spoken Dialogue Recognition and Its Evaluation on Operator's Speech at Call Centers. January 2006. Available online: https://pdfs.semanticscholar.org/9e9d/dcfbefd682979752fe4655ca39721c5fe211.pdf (accessed on 1 January 2021).
12. Pratap, V.; Sriram, A.; Tomasello, P.; Hannun, A.; Liptchinsky, V.; Synnaeve, G.; Collobert, R. Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters. 2007. Available online: arxiv.org/pdf/2007.03001.pdf (accessed on 8 July 2020).
13. Rebai, I.; BenAyed, Y.; Mahdi, W.; Lorré, J.P. Improving Speech Recognition Using Data Augmentation and Acoustic Model Fusion. *Procedia Comput. Sci.* **2017**, *112*, 316–322. [CrossRef]
14. Rustamov, S.; Akhundova, N.; Valizada, A. Automatic Speech Recognition in Taxi Call Service Systems. In *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering Emerging Technologies in Computing*; Springer: Cham, Switzerland, 2019; pp. 243–253. [CrossRef]
15. Sehgal, R.R.; Gaurav, R. Interactive Voice Response Using Automatic Speech Recognition Techniques for Call Centers. *Ssrn Electron. J.* **2018**. [CrossRef]
16. Seltzer, M.L.; Yu, D.; Wang, Y. An Investigation of Deep Neural Networks for Noise Robust Speech Recognition. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, UK, 26–31 May 2013. [CrossRef]
17. Soltau, H.; Liao, H.; Sak, H. *Neural Speech Recognizer: Acoustic-to-Word LSTM Model for Large Vocabulary Speech Recognition*; Interspeech: Shanghai, China, 2017. [CrossRef]
18. Wang, Y.; Zhang, C.; Gales, M.; Woodl, P.C. *Speaker Adaptation and Adaptive Training for Jointly Optimised Tandem Systems*; Interspeech: Shanghai, China, February 2018. [CrossRef]
19. Yadav, M.; Afshar, A. Speech Recognition: A Review. April 2018; Volume 6. Available online: https://www.researchgate.net/publication/324243764_Speech_Recognition_A_review (accessed on 1 January 2021).
20. Zhang, Y. Speech Recognition Using Deep Learning Algorithms. 2013. Available online: https://pdfs.semanticscholar.org/fbf6/2fad033af2c083bc3152066fd2cc4544da66.pdf?_ga=2.46858852.276772873.1571040186-2066189885.1571040186 (accessed on 1 January 2021).
21. Sound EXchange: HomePage. SoX. Available online: http://sox.sourceforge.net/ (accessed on 1 January 2021).
22. Yu, S.; Xu, C.; Liu, H. Zipf's Law in 50 Languages: Its Structural Pattern, Linguistic Interpretation, and Cognitive Motivation. July 2018. Available online: http://arxiv.org/abs/1807.01855 (accessed on 1 January 2021).
23. Aida-Zade, K.R.; Ardil, C.; Rustamov, S.S. Investigation of Combined Use of MFCC and LPC Features in Speech Recognition Systems. *World Acad. Sci. Eng. Technol.* **2006**, *13*, 74–80. Available online: https://www.researchgate.net/profile/Cemal_Ardil/publication/238744473_Investigation_of_Combined_use_of_MFCC_and_LPC_Features_in_Speech_Recognition_Systems/links/02e7e52dbbf5a80f2a000000/Investigation-of-Combined-use-of-MFCC-and-LPC-Features-in-Speech-Recognition-Systems.pdf (accessed on 1 January 2021).
24. Chodroff, E. Kaldi Tutorial. Eleanorchodroff.com. Available online: https://eleanorchodroff.com/tutorial/kaldi/index.html (accessed on 1 January 2021).
25. Veselý, K.; Ghoshal, A.; Burget, L.; Povey, D. *Sequence-Discriminative Training of Deep Neural Networks*; Interspeech: Shanghai, China, August 2013. Available online: https://doi.org/10.1007/978-3-319-64680-0_12 (accessed on 1 January 2021).
26. Miao, Y.; Zhang, H.; Metze, F. Speaker Adaptive Training of Deep Neural Network Acoustic Models Using I-Vectors. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 1938–1949. [CrossRef]
27. Scharenborg, O.; Ciannella, F.; Palaskar, S.; Black, A.; Metze, F.; Ondel, L.; Hasegawa-Johnson, M. Building an ASR System for a Low-Research Language Through the Adaptation of a High-Resource Language ASR System: Preliminary Results. In Proceedings of the International Conference on Natural Language, Signal and Processing, Taipei, Taiwan, 27 November–1 December 2017. Available online: https://pdfs.semanticscholar.org/d048/06f26584b923105c731e2e6f63b433dd96ea.pdf?_ga=2.39343587.1757005987.1594635569-419323592.1594635569 (accessed on 1 January 2021).
28. Jurafsky, D.; Martin, J.H. *Speech and Language Processing*, 2nd ed.; Pearson Education: London, UK, 2014; pp. 330–332. Available online: https://www.amazon.com/Speech-Language-Processing-Daniel-Jurafsky/dp/0131873210 (accessed on 1 January 2021).