


## Article

# Influence of Highly Inflected Word Forms and Acoustic Background on the Robustness of Automatic Speech Recognition for Human–Computer Interaction

Andrej Zgank 

Faculty of Electrical Engineering and Computer Science, University of Maribor, 2000 Maribor, Slovenia;  
andrej.zgank@um.si; Tel.: +386-2-220-7206

**Abstract:** Automatic speech recognition is essential for establishing natural communication with a human–computer interface. Speech recognition accuracy strongly depends on the complexity of language. Highly inflected word forms are a type of unit present in some languages. The acoustic background presents an additional important degradation factor influencing speech recognition accuracy. While the acoustic background has been studied extensively, the highly inflected word forms and their combined influence still present a major research challenge. Thus, a novel type of analysis is proposed, where a dedicated speech database comprised solely of highly inflected word forms is constructed and used for tests. Dedicated test sets with various acoustic backgrounds were generated and evaluated with the Slovenian UMB BN speech recognition system. The baseline word accuracy of 93.88% and 98.53% was reduced to as low as 23.58% and 15.14% for the various acoustic backgrounds. The analysis shows that the word accuracy degradation depends on and changes with the acoustic background type and level. The highly inflected word forms’ test sets without background decreased word accuracy from 93.3% to only 63.3% in the worst case. The impact of highly inflected word forms on speech recognition accuracy was reduced with the increased levels of acoustic background and was, in these cases, similar to the non-highly inflected test sets. The results indicate that alternative methods in constructing speech databases, particularly for low-resourced Slovenian language, could be beneficial.

**Keywords:** human–computer interaction; automatic speech recognition; acoustic modeling; highly inflected word forms; acoustic background

**MSC:** 68T10



**Citation:** Zgank, A. Influence of Highly Inflected Word Forms and Acoustic Background on the Robustness of Automatic Speech Recognition for Human–Computer Interaction. *Mathematics* **2022**, *10*, 711. <https://doi.org/10.3390/math10050711>

Academic Editors: Grigoreta-Sofia Cojocar and Adriana-Mihaela Guran

Received: 30 December 2021

Accepted: 22 February 2022

Published: 24 February 2022

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The convergence of Internet of Things (IoT) systems, services, and telecommunication networks has resulted in the omnipresence of human–computer interaction. Users can access services 24/7 by applying different devices. To support the human–computer interaction in a most natural way for users, procedures such as Automatic Speech Recognition (ASR) are needed [1]. A large vocabulary continuous speech recognition task can, on one side, be applied for standard human–computer interaction [2], or it can be used for producing text from various media material and user’s content [3]. The text can be used for navigating and controlling [4,5], subtitling, indexing [6], translating, topic detection [7] . . . Other examples of ASR applications besides human–computer interface input are the broadcast news speech recognition systems, massive open online courses, YouTube videos, and other various user content types, generated in intelligent ambient [8,9]. While an ASR can produce reasonable results in the case of typical conditions, the ASR’s accuracy degrades significantly for adverse acoustic conditions [10,11] or when more complex languages are adopted [12,13]. Various acoustic backgrounds increased significantly in the last decade, where users interact with devices in different situations and record content in

diverse environments. The omnipresent availability of smartphones in society has changed the role of who is recording and publishing the content.

The Slovenian University of Maribor (UMB) Broadcast News (BN) speech recognition system [14] is one such system for recognizing speech in a complex language with different acoustic backgrounds. It achieves approximately 15% worse overall speech recognition performance than similar systems for Western European languages. Possible causes are the linguistic characteristics of the Slovenian language, which is highly inflected and has a relatively free word order. An additional cause is the smaller amount of available Slovenian spoken language resources compared to other major languages like English, Mandarin, or German. While the research topic of under-resourced languages is covered broadly [12], there is far less attention given to the topic of highly inflected languages in automatic speech recognition. The preliminary analysis of achieved speech recognition results showed [15] that the worse performance belongs to the test set with the speech in the presence of an acoustical background, where the decrease can be even more than 40%. When similar English Broadcast News speech recognition systems were confronted with such degraded speech, the overall decrease of performance was in the region of 20% [16].

The decrease in the Slovenian speech recognition performance was the motivation to carry out a detailed analysis of how highly inflected word forms and acoustic background influence the speech recognition accuracy for human–computer interaction. We propose a novel analysis approach, where instead of using the general speech database, a separate, new type of speech database is constructed for tests. The new speech database is focused solely on highly inflected word forms originating from the same stem. This is needed to be able to perform a narrow analysis of specific language characteristics. The experiments are carried out with the Slovenian UMB Broadcast News speech recognition system. A particular emphasis is given to highly inflected word forms, where first, their effect on acoustic modeling is analyzed, and second, the combined influence of highly inflected word forms and acoustic background is studied. To the best of the author’s knowledge, no similar research on the combined influence of acoustic background and highly inflected word forms on the level of acoustic modeling, as proposed in this paper, has been carried out. Moreover, this presents a first attempt to construct a focused, highly inflected speech database to assess this phenomenon of human–computer interaction.

Our view is that such analysis can produce a detailed insight into how strongly these conditions lower speech recognition accuracy. The research hypothesis is that combined influence of acoustic background and highly inflected word forms intensify the speech recognition accuracy decrease and that it is possible to identify approximate thresholds, where new digital signal processing techniques and dedicated algorithms for training acoustic models for highly inflected languages could be beneficial to reduce this degradation. The proposed novel analysis is carried out for the Slovenian language, but the same approach with a dedicated speech database could also be used for other morphologically complex languages. Additional anticipation is that the analysis results could point toward efficient construction of speech databases for complex low-resourced Slovenian language, where it is challenging to collect several 1000 h of transcribed speech, as it is today’s standard for major languages.

The paper is organized as follows. First, the related work about the acoustic background and highly inflected languages for automatic speech recognition as part of human–computer interaction is presented. Different modules and functional aspects of an automatic speech recognition system, important for our research, are described in Section 3. The results and discussion are given in Section 4. The paper ends with the conclusions in Section 5.

## 2. Related Work

The robustness of speech recognition systems against the acoustic background is a mature research question, where the first research was coexistent with the fundamental automatic speech recognition research [17,18]. Despite the advancement of the automatic speech recognition research field, the question of robustness against the acoustic back-

ground, most frequently against noise, still presents a challenge, mainly in the area of developing new and improved methods of increasing the automatic speech recognition accuracy [19,20]. The question of robust automatic speech recognition can be addressed from two main viewpoints. The first one is focused on the reduction of acoustic background and improved feature extraction algorithms [21,22]. In the area of automatic speech recognition, Mel-frequency cepstral coefficients (MFCC) [23] preset a general feature extraction algorithm, which is used frequently. Different authors [24–26] proposed several other more complex feature extraction algorithms to improve the robustness. One of the frameworks, which includes a large number of various feature extraction algorithms is openSMILE [27], popular in the case of emotion recognition. More details about advanced feature extraction for speech recognition can be found in [28]. The second robustness viewpoint concentrates on robust acoustic modeling approaches, where neural networks have mainly been used in recent years [29]. One of the frequently used solutions how to improve the robustness of neural networks regarding the acoustic background is data augmentation [30]. In this case, the original acoustic training data is extended with artificially generated copies, where the various acoustic backgrounds are added [31,32]. Moreover, other original audio data characteristics, like speed or pitch [33] can be altered to augment the training set and make it more robust to different conditions. The area of highly inflected languages for automatic speech recognition is a much less addressed topic. This is mainly due to the fact that the majority of currently commercially interesting languages do not belong to this category or are compensating for the limitations of a highly inflected language with a large amount of available spoken language resources. In the case of highly inflected languages, the language modeling question was given more importance [34], but the question of how highly inflected word forms influence acoustic modeling is a much less researched area. One of the first research works focused on the effects of highly inflected word forms for language modeling for automatic speech recognition was carried out for the Czech [35,36] and Slovenian languages [37]. The most frequently proposed solution for addressing the question of high inflection on the level of language modeling is to apply a non-word basic unit for modeling [38]. The idea behind this approach is to split the highly inflected word forms into shorter forms, which have a higher frequency in spoken language resources. In general, as a result of this approach, the out-of-vocabulary rate is usually reduced significantly, which might lead to improved speech recognition accuracy. The first disadvantage of this modeling approach is the reduced length of basic units, which increases acoustic confusability. The second disadvantage is the reduced modeling power of the language model, as the sequence of n-grams now estimates shorter linguistic parts of a sentence. Different types of basic units for language modeling and their combinations were proposed: Morphological [39,40], grammatical [41,42], stem-ending [43]. The proposed approaches improve the speech recognition accuracy to some extent, but the lag to major language performance can still be observed. In recent years, similar approaches used for highly inflected language modeling for automatic speech recognition were also adapted successfully to machine translation [38].

On the level of acoustic modeling for highly inflected automatic speech recognition, less research work was presented in the literature [44–46]. The shorter basic units from language modeling were incorporated directly into acoustic modeling, with necessary modifications to the decoding algorithms [47], which was used partly to compensate the reduced estimation power of the shorter n-gram units.

### 3. Materials and Methods

The performance of a speech recognition system depends largely on the acoustic characteristics of the input signal and on the properties of the supported language. Both conditions, as well as the experimental setup will be described below.

### 3.1. Acoustic Background and Speech Recognition

If the input signal  $x(t)$  contains only speech signal  $s(t)$  and acoustic background conditions are clean, high word accuracy can be expected, even for complex tasks such as continuous or spontaneous speech recognition [1]. In the case when the acoustic background signal  $b(t)$  is present, the speech signal gets corrupted, and the word accuracy decreases. The corrupted input signal  $x'(t)$  can be defined as in Equation (1):

$$x'(t) = s(t) + b(t) \quad (1)$$

The word accuracy decrease depends on the type and characteristics of the background signal  $b(t)$  (e.g., noise, music, speech . . . ) and its overall energy and spectral characteristics in comparison with the main speech signal  $s(t)$ . The energy ratio between both signals can be estimated with the signal-to-noise ratio (SNR, Equation (2)), where the background signal is treated as noise:

$$\text{SNR(dB)} = 10 \cdot \log \frac{P_{\text{speech}}}{P_{\text{background}}} \quad (2)$$

where  $P_{\text{speech}}$  denotes the overall energy of the speech signal and  $P_{\text{background}}$  denotes the overall energy of the background signal. The most important characteristic of background signal that influences the level of corruption is the spectrum, which varies, and is non-stationary.

In the detailed analysis presented in this paper, the acoustic models were trained on a speech database with various acoustic backgrounds. The baseline test set used only clean speech, while the dedicated test sets applied music as background, which corrupted the original clean speech signal. Music as a background source presents a realistic scenario, as such type of degradation occurs in the case of user content and input signals captured in human–computer interaction in intelligent ambient.

### 3.2. Acoustic Modeling and Highly Inflected Language

There are approximately 7000 live languages in the world, with 90% of them having less than 100,000 speakers. The human language technology support for various languages is important for their existence in the digital era, as it provides the core technology for natural human–computer interaction. Different languages can present a challenging task for a speech recognition system, due to their properties. Examples of such languages are tonal, highly inflected, or agglutinative languages. In the case of a highly inflected language, a single word stem is modified with many different suffixes or prefixes, which results in a large number of word forms. For some cases, the resulting number of word forms, modified from common stem, can exceed 100. As a consequence, an automatic speech recognizer's lexicon has to include a much larger number of words to achieve the same out of vocabulary rate as those speech recognition systems for a non-inflected language. It is crucial to achieving low out of vocabulary rate, as all the missing words from the vocabulary are being reflected directly in speech recognition errors as substitutions, with additional errors (approx. factor  $1.5\times$  might apply) coming from language modeling. Highly inflected language modeling approaches, such as stem-ending ones, can, to some extent, limit this effect of high out of vocabulary rate, but they lose a part of the n-gram prediction power, as shorter basic units span over shorter linguistic structures of a sentence.

Some of the languages belonging to the highly inflected group are Arabic, Slavic (e.g., Russian, Czech, Slovak, Polish, Slovenian, ...), Baltic (e.g., Lithuanian, and Latvian), and Indic languages. Examples of the Slovenian highly inflected word “hiša” (ENG: “house”) in different cases for singular, dual, and plural, are given in Table 1.

Highly inflected word forms can be very similar acoustically, as can be seen from the example in Table 1, where only the last (or last few) phoneme of a relatively short word is modified. This acoustic confusability can to some degree be compensated with the n-grams of a statistical language model but can still present a severe problem for the speech decoder, especially in the case of the presence of acoustic background or accented speech. The Slovenian language has more than 50 dialects, which are spoken by approximately

2 million speakers. This results in a vast acoustic diversity, making the Slovenian language very suitable for our experiments of developing a dedicated speech database with highly inflected word forms. An additional challenge is given by the speaking style type, where two main categories are read/planned and spontaneous speech. In the case of spontaneous speech, an end vowel reduction often occurs in many Slovenian dialects, which additionally increases the acoustic confusability of highly inflected word forms. These characteristics clearly present the complexity of the automatic speech recognition task in the case of such languages.

**Table 1.** Example of Slovenian highly inflected word “hiša”.

Singular		Dual		Plural	
Word	Suffix	Word	Suffix	Word	Suffix
hiša	-a	hiši	-i	hiše	-e
hiše	-e	hiš	/	hiš	/
hiši	-i	hišama	-ama	hišam	-am
hišo	-o	hiši	-i	hiše	-e
hiši	-i	hišah	-ah	hišah	-ah
hišo	-o	hišama	-ama	hišami	-ami

### 3.3. Speech Databases

The Slovenian BNSI Broadcast News speech database [48] belongs to the group of similar speech databases for Central and Eastern European languages [3,49], which were developed in the last two decades with the objective to provide suitable speech recognition resources for complex under-resourced languages. The BNSI speech database was used for research in the area of spoken language technologies and the design of the UMB BN speech recognition system. The Slovenian BNSI Broadcast News speech database is available via European Language Resources Association (ELRA/ELDA). Due to the specific requirements of the proposed analysis procedure, where only the acoustic characteristics should be considered, two additional speech databases were needed for the evaluation. The first one was the Slovenian SNABI SSQ Studio speech database, which was used for evaluating the general influence of acoustic background on continuous speech recognition accuracy. The Slovenian SNABI SSQ Studio speech database is available via Clarin SI. The second one was a dedicated test speech database, HI\_SI, built for evaluating the influence of highly inflected word forms on acoustic modeling. The test scenarios, with isolated words, were selected for the evaluation, to avoid the influence of the language model, which could mask the acoustic differences which needed to be analyzed.

The Slovenian BNSI Broadcast News speech database consists of a total of 36 h of spoken material, where 30 h are used for acoustic training, 3 h for development, and the additional 3 h for evaluation. There are 1565 different speakers, 1069 of them are male and 477 are female. The gender of the remaining 19 speakers was unknown.

An important factor influencing speech recognition accuracy is acoustic background. There are several possible background types present [50] in a real-life environment: noise [10], speech, music. In a broadcast news speech database, the acoustic background is reflected in the parameter called acoustic conditions suitably simulating different conditions in which human–computer interaction needs to operate. The acoustic conditions and acoustic background in the BNSI speech database are presented in Table 2.

A large amount of the various acoustic conditions in the BNSI database is suitable for acoustic training, as it helps to increase the robustness of acoustic models against different acoustic conditions present in the evaluation set. However, it must also be taken into account that it could be difficult to train acoustic models for those acoustic conditions reliably with a relatively small proportion of the training set (e.g., F2, F5, FX).



**Table 2.** Acoustic conditions in the Slovenian BNSI Broadcast News speech database.

Acoustic Condition	Speech Type	Acoustic Background	Ratio (%)
F0	read	clean	36.56
F1	spontaneous	clean	16.23
F2	read/spontaneous	telephone channel	1.65
F3	read/spontaneous	music	6.02
F4	read/spontaneous	other	37.63
F5	nonnative	various	0.05
FX	other	other	1.86

The SNABI SSQ Studio test set used for the first part of the analysis consists of recordings of 37 speakers, where each speaker uttered different words in a silent studio environment. Thus, the baseline acoustic conditions of the test set were comparable to the ones present in the F0 and F1 acoustic condition parts of the training set. The first test set was based on isolated digits (12 words in the vocabulary, 441 recordings). The isolated digits' scenario is a non-complex one, but it is used frequently in different languages and applications, and thus enables a wide overall comparison of results. The second one, on city names (37 words in the vocabulary, 1360 recordings), is a more acoustically diverse test set, where the city names guarantee the absence of highly inflected word forms. This setup was used for baseline evaluation in the first part of the analysis. Dedicated test sets, where the baseline test set was combined with various acoustic backgrounds, were prepared for our experiment. The first type of acoustic background was instrumental music, similar to those frequently present in different media. The second type of acoustic background was based on songs, where instrumental music was combined with singing. This acoustic background type will be denoted in the remaining part of this paper as vocal music. Different types and SNR ratios of acoustic background generated for the dedicated SNABI test scenarios are presented in Table 3.

**Table 3.** Acoustic background in the dedicated SNABI test sets.

Acoustic Background Level	Instrumental Music SNR(dB)	Vocal Music SNR(dB)
Low	30.74	30.14
mid-low	22.30	22.06
mid	15.15	15.03
mid-high	10.77	10.69
high	5.10	5.03

The first part of the evaluation in Section 4 was carried out with the baseline clean and five acoustically degraded dedicated test sets. In the case of broadcast news recordings, the majority of the acoustic background has lower energy levels (i.e., higher SNR), and its type is usually instrumental.

The proposed HI\_SI speech database was built for providing the dedicated test set with highly inflected word forms in a silent environment and in the presence of an acoustic background. During the preparation of experiments, several existing Slovenian speech databases were taken into account, but none of them comprised a test set suitable for evaluating acoustic confusability between highly inflected word forms. The HI\_SI speech database contains recordings of 10 speakers, uttering various test sets/scenarios with six different highly inflected words generated from a single stem. In addition, one reference test scenario was recorded, which contains only words with a different stem in the basic word form. Three highly inflected test scenarios and the reference basic word form scenario were then selected as a baseline for the second evaluation part:

- H test set: Slovenian word “hiša” (Eng.: house), 6 inflected word forms,
- T test set: Slovenian word “telefon” (Eng.: telephone), 6 inflected word forms,
- M test set: Slovenian word “monitor” (Eng.: monitor), 6 inflected word forms,
- Reference test set: 6 acoustically diverse Slovenian words.

Instrumental music at three different levels was applied as an acoustic background for the second part of the evaluation. Different SNR ratios of acoustic background generated for the selected HI\_SI test scenarios are presented in Table 4.

**Table 4.** Acoustic background in the dedicated HI\_SI test sets.

Acoustic Background Level	H Test Set SNR (dB)	T Test Set SNR (dB)	M Test Set SNR (dB)	Reference Test Set SNR (dB)
low	31.18	30.64	30.96	31.02
mid	14.86	15.17	14.92	15.21
high	5.47	5.29	5.18	5.20

The acoustic properties of the HI\_SI test sets, combined with the instrumental music as background, were comparable with the SNABI SSQ test sets, which is a favorable starting point for analyzing the influence of highly inflected word forms on acoustic modeling in the presence of an acoustic background.

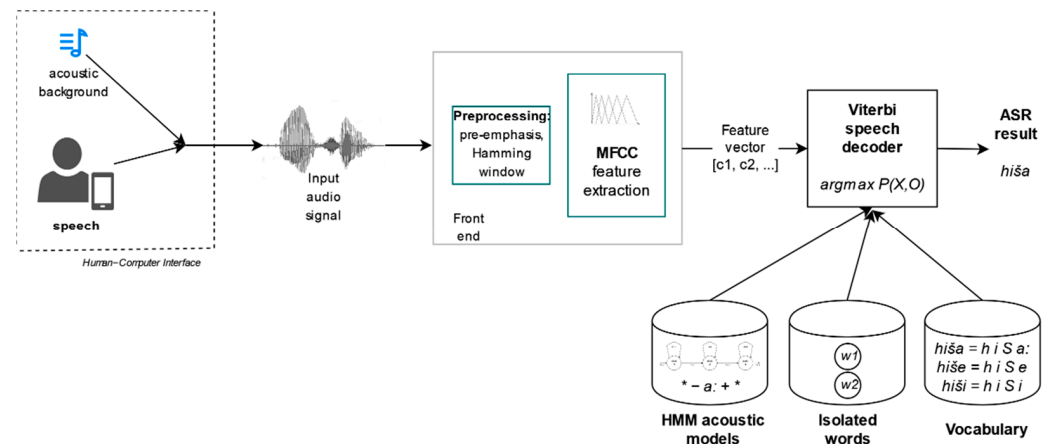
The highly inflected test scenario was prepared in such a way that the city names isolated words’ grammar was taken as a baseline and augmented with the highly inflected word forms from the particular HI\_SI test set. Four additional isolated grammars were produced as an end result.

### 3.4. Experimental Setup

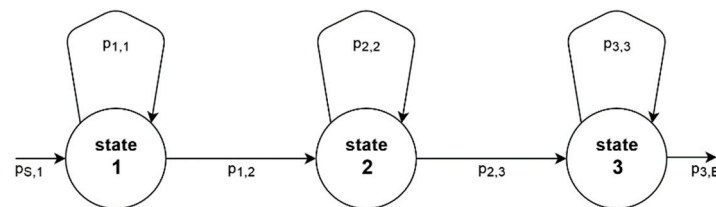
A general automatic speech recognition system is comprised of five basic building blocks (Figure 1; from left to right): Feature extraction front-end, acoustic models, language model, lexicon, and decoder. Front-end first pre-processes the input speech signal including the pre-emphasis and windowing and then extracts features. The extracted feature vectors are used as the input data for the speech decoder, which is applied to estimate the most probable hypothesis according to the models and feature vectors. There are three different types of models needed by a general speech decoder. The first are acoustic models, which represent the acoustic-phonetic parameters of acoustic material involved in the training process [50]. The language model is used to estimate the language characteristics in the form of a statistical model or grammar. It is built on a large text corpus in form of n-grams or using a grammatical description language in case of isolated words grammar. The last model is the phonetic vocabulary, which is used to transform the orthographic representation of words into phoneme form, which is generally needed for acoustic modeling and recognition. The output of a speech decoder is text result, which represents the word sequence uttered by the speaker in the input signal. A scheme of an automatic speech recognition system, with all the test sets included in the proposed evaluation, is given in Figure 1. The evaluation is divided into first and second part. The first one is focusing on various acoustic background, while the second one analyses the influence of highly inflected word forms with the proposed approach.

The experimental setup was designed in such a way that acoustic-phonetic properties played the most important role in the design and analysis process. The focus was given to the acoustic models, whose training procedure is described below. The automatic speech recognition system used for the analysis was built on three state left-right hidden Markov models (HMM), which were applied for acoustic modeling (Figure 2).

The limited amount of available Slovenian speech training data dictated the decision to use the HMM acoustic models instead of deep neural network (DNN) models. The first acoustic models’ type usually performs more stable, particularly under challenging conditions.



**Figure 1.** Scheme of the automatic speech recognition system used for the analysis.



**Figure 2.** Three state left-right HMM acoustic model.

### 3.4.1. Preprocessing the Speech Signal

The captured speech signal  $s$  must be first prepared for acoustic training using the preprocessing methods. The pre-emphasis, as defined in Equation (3):

$$s'[n] = s[n] - 0.97 \cdot s[n-1] \quad (3)$$

was applied to amplify the higher frequency band of the speech signal  $s$  with a high pass filter. Then, the signal  $s$  was split to produce frames, which were 25 ms long and were shifted with 10 ms step. A Hamming window function, defined in Equation (4):

$$s'[n] = \left\{ 0.54 - 0.46 \cos\left(\frac{2\pi(n-1)}{N-1}\right) \right\} \cdot s[n] \quad (4)$$

was used to transform samples in each frame.  $N$  represented the length of Hamming window and  $n$  the particular sample.

### 3.4.2. Feature Extraction

After preprocessing, the feature extraction was done. Twelve Mel-scale frequency cepstrum coefficients (MFCC) were used as basic feature vectors. The MFCC were built around Mel-scale filter bank, which represented the nature of non-linear human speech perception and was linking the Mel ( $m$ ) and frequency ( $f$ ) dimension, as defined in Equation (5):

$$m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right). \quad (5)$$

The particular MFCC coefficient  $c_i$  was calculated using the Equation (6):

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^N m_j \cos\left(\frac{\pi i}{N}(j-0.5)\right). \quad (6)$$



As the 13th feature, the signal energy  $E$  was added to the vector, as defined in Equation (7):

$$E = \log \sum_{n=1}^N s^2[n]. \quad (7)$$

The basic set of 13 features was needed to calculate the first and second-order derivatives, which proved to model the short-term variations present in the speech signal. The final feature vector involved in the acoustic modeling contained 39 different coefficients extracted from the input speech signal.

### 3.4.3. The Acoustic Model and Training

The speech recognizer's acoustic HMM models applied a three-state topology with a weighted sum of continuous Gaussian probability density functions (PDF). The acoustic model  $\lambda$  was set as (Equation (8)):

$$\lambda = (A, B, \pi), \quad (8)$$

where  $A$  presented the state transition probabilities ( $a_{ij}$ ),  $B$  the observation probability distribution ( $b_j$ ) per each state  $j$  and  $\pi$  the initial values. For each observation  $o_t$  the probability  $b_j$  that it was generated, was defined as (Equation (9)):

$$b_j(o_t) = \sum_{m=1}^M c_{jm} \mathcal{N}(o_t; \mu_{jm}, \Sigma_{jm}), \quad (9)$$

where  $c_{jm}$  was the weight of each mixture,  $M$  the total number of mixtures and  $\mathcal{N}$  the Gaussian PDF with mean vector  $\mu$  and diagonal covariance matrix  $\Sigma$  (Equation (10)):

$$\mathcal{N}(o; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(o-\mu)'\Sigma^{-1}(o-\mu)} \quad (10)$$

The acoustic models were speaker and gender independent, with graphemes as basic acoustic units. Graphemes were used instead of phonemes as basic units as they reduce the influence of phonetic vocabulary errors which result from the complex grapheme-to-phoneme task for the Slovenian language. The HMM acoustic models were trained using the Baum-Welch re-estimation, which is defined in the form of forward probability  $\alpha_j$  (Equation (11)) using the forward recursion:

$$\alpha_j(t) = \left[ \sum_{i=2}^{N-1} \alpha_i(t-1) a_{ij} \right] b_j(o_t), \quad (11)$$

where  $\alpha_1(1) = 1$  and  $\alpha_j(1) = a_{1j} b_j(o_1)$  are set as initial conditions and  $j$  is  $1 < j < N$ . The final step is defined as (Equation (12)):

$$\alpha_N(T) = \left[ \sum_{i=2}^{N-1} \alpha_i(T) a_{iN} \right] \quad (12)$$

The backward probability  $\beta_j(t)$  is defined by Equation (13) using the backward recursion:

$$\beta_i(t) = \sum_{j=2}^{N-1} a_{ij} b_j(o_{t+1}) \beta_j(t+1), \quad (13)$$

where  $i$  and  $t$  are  $1 < i < N$  and  $T > t > 1$ . The initial value is set as  $\beta_i(T) = a_{iN}$  with the final step defined as (Equation (14)):

$$\beta_1(1) = \sum_{j=2}^{N-1} a_{1j} b_j(o_1) \beta_j(1). \quad (14)$$

### 3.4.4. Training of Context-Independent Acoustic Models

The above-defined algorithms were used for training the acoustic models in three steps. The first two used context-independent modeling. The last one also modeled the

context of a single acoustic unit, which can improve the acoustic modeling performance significantly [51]. This approach was applied to improve the quality of speech database transcriptions stepwise. The original transcriptions were produced manually during the speech database development but contained some errors.

First, the context-independent acoustic models' initial parameters were set as global values, equal for all different models. The initial values were estimated using the Baum-Welch re-estimation. For acoustic training, a small subset was randomly selected from the full training set. Three iterations of the training procedure were carried out. The results were acoustic models with 1 Gaussian PDF per state. Next, the forced realigning of the speech database's transcriptions was used to detect the outliers in the training set.

In the second step, the context-independent acoustic models were initialized from the scratch. The initial acoustic models with 1 Gaussian PDF were trained in a stepwise manner until reaching the mixture of 32 Gaussian PDF per state. Such a type of acoustic model has a better generalization effect. They are more suitable for the evaluation of an unseen test set comprising different acoustic conditions. These acoustic models were again applied for the forced-realigning procedure. The result was an improved version of training set speech transcriptions.

#### 3.4.5. Training of Context-Dependent Acoustic Models

The third step started again with the initialization of acoustic models. This time local initial values for each acoustic unit were used. The cross-word context-dependent (trigrapheme) acoustic models were created after seven training iterations. The context-dependent acoustic models improve the performance in presence of the coarticulation effect [52]. The context-dependent acoustic models have a large number of free parameters. These were reduced with the use of a decision tree-based clustering algorithm. Its objective was to tie together all similar acoustic models' states. A data-driven approach [53] based on the acoustic models' confusion matrix was used to produce broad classes. In the end, the number of Gaussian PDF per state was increased to 16. These were the final acoustic models, which were applied to the speech recognition task. For more details on the UMB BN speech recognition system see [14,15].

### 4. Results and Discussion

The first part of the analysis was focused on the influence of acoustic background on speech recognition accuracy. The experiments were carried out in several steps using the UMB BN speech recognition system and dedicated test sets generated from the SNABI SSQ Studio database. The speech recognition results for isolated words recognition are presented in the form of Word Accuracy (WA), which is defined as in Equation (15):

$$\text{Word accuracy}(\%) = \frac{H - I}{N} \cdot 100 \quad (15)$$

where  $H$  denotes the number of correctly recognized words in the test set,  $I$  is the number of insertions and  $N$  denotes the number of all words in the test set.

For comparison reasons, a short overview of previous general speech recognition results with the UMB BN large vocabulary continuous speech recognition system will be given. The comparable setup of the UMB BN system, with the BNSI Broadcast News speech database [4], achieves a 73.26% word accuracy on the reference BNSI evaluation set, which includes all acoustic conditions. In the case of clean read studio speech [14], the word accuracy was 84.81%, but it dropped to just 63.88% when various acoustic backgrounds (F3, F4, and Fx conditions) were also present. These results can be directly compared with the Slovenian ASR system developed by Golik et al. [54], who has used the identical BNSI Broadcast News speech database, but different system architecture achieved 72.2% word accuracy on the evaluation set. Additional comparisons can be made with broadcast news speech recognition systems for other languages or with cross-lingual setups [55,56]. The broadcast news ASR system for Slovak, a similar Slavic language, presented by Vizslay et al. [57],

performed with 78.04% word accuracy on a more extensive training speech database. The Slovenian ASR system [58] built with cross-lingual bootstrapped acoustic models and re-trained on lightly supervised public data achieved 78.51% word accuracy. The comparison with published results shows that the UMB BN large vocabulary continuous speech recognition system has similar performance and can be used to carry out the proposed analysis.

First, the analysis was devoted to evaluating both baseline SNABI test sets without acoustic background with the UMB Broadcast News speech recognition system baseline. Word accuracy for isolated digits was 93.88% and for the city names 98.53%. Although the city names test set includes more words in the grammar than the isolated digits' test set, it achieved better performance, due mainly to the longer words. In the case of isolated digits, the average word length was 3.7 graphemes, and for the city names the average word length was 5.2 graphemes. Longer words are acoustically diverse and thus give the speech recognition decoder more data to estimate the most probable grapheme sequence.

In the second step of the first part of the analysis, the SNABI isolated digits and city names' test sets with mixed acoustic backgrounds were tested with the UMB Broadcast News speech recognition system. The first test set applied instrumental music for the acoustic background and the second one used vocal music. The speech recognition results for five different acoustic background levels per test set are given in Tables 5 and 6.

**Table 5.** The isolated digits' speech recognition results in the presence of various acoustic backgrounds.

Test Set	Acoustic Background Level	Instrumental Music WA (%)	Vocal Music WA (%)
isolated digits	clean	93.88	93.88
isolated digits	low	88.21	84.35
isolated digits	mid-low	69.16	65.08
isolated digits	mid	42.86	42.86
isolated digits	mid-high	29.25	33.56
isolated digits	high	23.58	28.12
isolated digits	average background level	50.61	50.80

**Table 6.** The city names speech recognition results in the presence of various acoustic backgrounds.

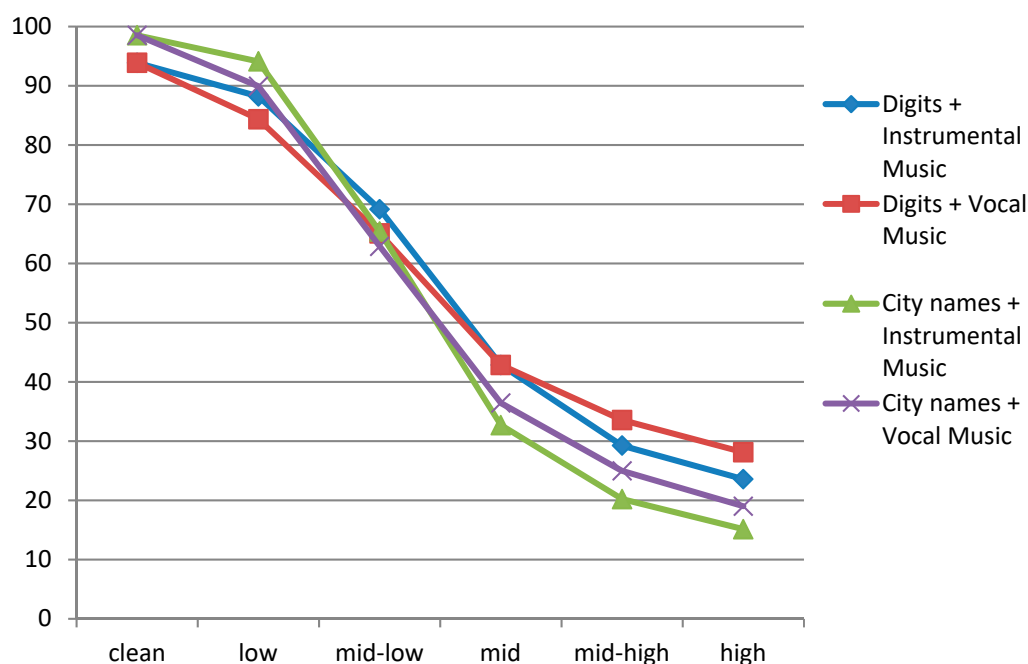
Test Set	Acoustic Background Level	Instrumental Music WA (%)	Vocal Music WA (%)
city names	clean	98.53	98.53
city names	low	94.12	89.93
city names	mid-low	65.39	62.89
city names	mid	32.70	36.44
city names	mid-high	20.21	24.98
city names	high	15.14	19.03
city names	average background level	45.51	46.65

When isolated digits mixed with acoustic background were recognized, the instrumental music decreased word accuracy from 93.88% to 23.58%. In the case of low and mid-low acoustic background levels, word accuracy was 88.21% and 69.16%, respectively. In the case when the vocal music was used for the acoustic background, the word accuracy reduced from 93.88% to 28.12%. Word accuracy of 84.35% and 65.08% was achieved with the low and mid-low acoustic background levels. The average word accuracy with instrumental music as a background was 50.61%, and 50.80% with vocal music as a background.

These results show that the type of acoustic background does have an influence on speech recognition accuracy. In the case of lower background level, the vocal music produced higher degradation than the instrumental music. This can be explained by the effect where the vocal part of music decreases the robustness of acoustic models, due to its similarity with ordinary speech found in uttered sentences.

An instrumental music background decreased the word accuracy of the city names test set from 98.53% to 15.14%, respectively. At low and mid-low background levels, the word accuracy was 94.12% and 65.39%, respectively. The vocal music as acoustic background degraded the word accuracy from 98.53% to 19.03%, respectively. The city names' average word accuracy was 45.51% for instrumental music background and 46.65% for the vocal music background. In the case of the city names test scenario, the difference in average word accuracy between both types of the acoustic background was slightly higher than for the previous experiment with isolated digits.

A detailed comparison of all four various test scenarios included in the first part of the analysis is given in Figure 3. The comparison between test scenarios confirms that the influence of acoustic background level also depends on the type and complexity of the test scenario. In the case of the city names test scenario, which has more words in the speech recognizer's vocabulary than the isolated digits scenario, the word accuracy decrease is smaller for low-level acoustic background (low condition), but the word accuracy degradation increased for higher levels of acoustic background (mid-high and high conditions). A plausible explanation is that the acoustic background increased the acoustic confusability between words in the speech recognizer's search space significantly.



**Figure 3.** Comparison of different acoustic background levels on the speech recognizer's word accuracy.

The influence of acoustic background type on speech recognition accuracy swaps with the level of acoustic background. While for lower acoustic background levels vocal music degraded speech recognition performance in a more severe way, the situation changed after introduction of the medium level of acoustic background. In the case of higher acoustic background level, the instrumental music showed higher degrading factors on speech recognition accuracy. A possible reason is that vocals in music at the low level of acoustic background present a sort of background speech, which is known to decrease the word accuracy significantly. With the increased level of acoustic background, this effect of

background speech diminished, and the degradation of musical acoustic background as a whole began to be the most important influence factor.

The second part of the analysis was focused on speech recognition accuracy for highly inflected words. Utterances without acoustic background were tested first, with the goal to define the baseline. In the next step, recordings of highly inflected words were evaluated, with various levels of acoustic background (Table 7).

**Table 7.** Speech recognition results for the highly inflected HI\_SI test sets.

Acoustic Background Level	H Test Set WA (%)	T Test Set WA (%)	M Test Set WA (%)	Non-Highly Inflected Test Set WA (%)
clean	71.7	63.3	80.0	93.3
low	76.7	60.0	76.7	93.3
mid	40.0	28.3	43.3	41.7
high	15.0	11.7	13.3	13.3
average	43.9	33.3	44.4	49.4

The reference test set (Table 7) comprised only non-highly inflected words and city names and was used as a control case, showing how the acoustic background influenced speech recognition accuracy for this specific scenario. The achieved results (WA between 93.3% and 13.3%) with this set were comparable to the ones in the first part of the analysis. This confirms that the newly proposed highly inflected HI\_SI speech database provides comparable data for evaluation. The highly inflected test sets produced significantly worse results when tested in clean acoustic conditions. The word accuracy dropped to 71.7%, 63.3%, and 80.0% for the H, T, and M test sets, respectively. The detailed analysis of speech recognition results revealed that, in the case of a clean acoustic background, all speech recognition errors occurred only between the highly inflected word forms (e.g., uttered word form “monitorja” was, for example, substituted with the form “monitorju”). The most frequent highly inflected word form substitution occurred in the H test set where, in 47.1% of error cases, the uttered word form “hiša” was misrecognized as word form “hišah”. In the case of both highly inflected words, the difference in suffix is only minimal—the first suffix “-a” is acoustically very similar to the second suffix “-ah”. This ending phoneme /h/ can frequently be reduced in various Slovenian accents. Moreover, the phoneme /h/ acoustic characteristics (spectral energy, time variance) can present a difficult task for a speech recognition system. This type of substitution errors and a significant decrease in word accuracy point-out clearly the complexity of speech recognition for highly inflected languages. When a low-level acoustic background was added to the highly inflected HI\_SI test sets T and M, additional word accuracy degradation was observed—60.0% and 76.7%, respectively. The first substitutions to the city names were observed in this case. This shows that the acoustic background increased the acoustic confusability additionally, beyond the point of modified word endings. For the Reference (93.3%) and H test set (76.7%), the accuracy was preserved or even improved. The possible cause for this is the statistical approach used for acoustic modeling. The decrease of word accuracy between the Reference test set and highly inflected H, T, and M, test sets was still comparable with the clean acoustic background configuration. With the increased level of acoustic background (mid and high levels), the word accuracy decreased statistically significantly for all four test sets similarly. The effect of acoustic confusability of highly inflected word forms was now minimized, as the higher level of the acoustic background masked the speech signal and its characteristics important for speech recognition algorithms. This was also confirmed with the detailed analysis of speech recognition results, where the substitutions were almost evenly distributed between errors in highly inflected word forms and city names.

The overall results show that, in the most frequent speech recognizers’ operating conditions, the additional absolute word accuracy degradation between 15% and 25% caused

by acoustic modeling can be expected for highly inflected languages. The experimental results confirmed the research hypothesis. A specific area above 15 dB SNR was identified from results where it would be of particular benefit to apply advanced acoustic modeling methods for highly inflected languages to improve the speech recognition results. Below the 15 dB SNR level, the influence of highly inflected word forms on speech recognition accuracy diminishes, and the general influence of acoustic background comes in front.

## 5. Conclusions

The proposed analysis approach showed the severe impact of acoustic background on speech recognition accuracy. While acoustic background at low energy decreased the word accuracy in a limited way, the high energy of acoustic background presented a limitation for using an ASR. One of the possible solutions would be to treat both cases separately, using different background reduction/decomposition methods.

The speech recognition analysis on highly inflected test scenarios showed a statistically significant additional impact of this language characteristic on word accuracy, particularly for none or a low level of acoustic backgrounds. This confirms that some languages are more challenging for automatic speech recognition tasks.

It would be interesting to extend the proposed analysis to other languages to check if similar estimations can also be observed for them. The prerequisite for such a task would be the availability of a dedicated highly inflected acoustic test set in the analyzed language. Currently, the well-established approach in acoustic modeling for automatic speech recognition is to collect as many hours of transcribed speech material as possible. This can result in several 10,000-h speech databases for major languages. However, this approach is unrealistic for many low-resource languages with a limited number of speakers and limited economic interest. The results of this analysis could in the future pave the way for an alternative acoustic modeling approach where the speech database would be designed in a way to better cover such language properties as are highly inflected word forms. This could reduce the need for collecting large amounts of transcribed speech material and ease the development.

**Funding:** The Slovenian Research Agency partially funded this research under Contract Number P2-0069 “Advanced methods of interaction in Telecommunication”.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The Slovenian SNABI SSQ Studio speech database is available via Clarin SI, <http://hdl.handle.net/11356/1051> (accessed on 29 December 2021). The Slovenian BNSI Broadcast News speech database is available via ELRA/ELDA, <http://www.islrn.org/resources/502-280-144-938-4/> (accessed on 29 December 2021).

**Acknowledgments:** The author thanks all those who contributed their speech recordings to the spoken language’s resources used in these experiments.

**Conflicts of Interest:** The author declares no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## References

1. Lee, C.H. On Automatic Speech Recognition at the Dawn of the 21st Century. *IEICE Trans. Inf. Syst.* **2003**, *E86-D*, 377–396.
2. Maskeliunas, R.; Ratkevicius, K.; Rudzionis, V. Voice-based Human-Machine Interaction Modeling for Automated Information Services. *Electron. Electr. Eng.* **2011**, *110*, 109–112. [\[CrossRef\]](#)
3. Pleva, M.; Juhar, J. Building of Broadcast News Database for Evaluation of the Automated Subtitling Service. *Commun.-Sci. Lett. Univ. Zilina* **2013**, *15*, 124–128. [\[CrossRef\]](#)
4. Miśkowska, M. Discriminant Analysis of Voice Commands in the Presence of an Unmanned Aerial Vehicle. *Information* **2021**, *12*, 23. [\[CrossRef\]](#)
5. Valizada, A.; Akhundova, N.; Rustamov, S. Development of Speech Recognition Systems in Emergency Call Centers. *Symmetry* **2021**, *13*, 634. [\[CrossRef\]](#)



6. Szaszak, G.; Tundik, A.M.; Vicsi, K. Automatic speech to text transformation of spontaneous job interviews on the HuComTech database. In Proceedings of the 2011 2nd International Conference on Cognitive Infocommunications (CogInfoCom), Budapest, Hungary, 7–9 July 2011.
7. Zlacký, D.; Staš, J.; Juhar, J.; Čížmar, A. Term weighting schemes for Slovak text document clustering. *J. Electr. Electron. Eng.* **2013**, *6*, 163–166.
8. Gondí, S.; Pratap, V. Performance Evaluation of Offline Speech Recognition on Edge Devices. *Electronics* **2021**, *10*, 2697. [\[CrossRef\]](#)
9. Beňo, L.; Pribiš, R.; Drahoš, P. Edge Container for Speech Recognition. *Electronics* **2021**, *10*, 2420. [\[CrossRef\]](#)
10. Pervaiz, A.; Hussain, F.; Israr, H.; Tahir, M.A.; Raja, F.R.; Baloch, N.K.; Ishmanov, F.; Bin Zikria, Y. Incorporating Noise Robustness in Speech Command Recognition by Noise Augmentation of Training Data. *Sensors* **2020**, *20*, 2326. [\[CrossRef\]](#)
11. Gnanamanickam, J.; Natarajan, Y.; Sri, S.P.K. A Hybrid Speech Enhancement Algorithm for Voice Assistance Application. *Sensors* **2021**, *21*, 7025. [\[CrossRef\]](#)
12. Besacier, L.; Barnard, E.; Karpov, A.; Schultz, T. Automatic speech recognition for under-resourced languages: A survey. *Speech Commun.* **2014**, *56*, 85–100. [\[CrossRef\]](#)
13. Wołk, K.; Wołk, A.; Wnuk, D.; Grześ, T.; Skubis, I. Survey on dialogue systems including slavic languages. *Neurocomputing* **2021**, *477*, 62–84. [\[CrossRef\]](#)
14. Maučec, M.S.; Žgank, A. Speech recognition system of Slovenian broadcast news. In *Speech Technologies*; InTech: Rijeka, Croatia, 2011; pp. 221–236.
15. Gank, A.; Donaj, G.; Maučec, M.S. UMB Broadcast News 2014 continuous speech recognition system: What is the influence of language resources' size? Language technologies. In Proceedings of the 17th International Multiconference Information Society—IS 2014, Ljubljana, Slovenia, 9–10 October 2014; Volume G.
16. Raj, B.; Parikh, V.; Stern, R. The effects of background music on speech recognition accuracy. In Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, Munich, Germany, 21–24 April 1997; pp. 851–854. [\[CrossRef\]](#)
17. Gong, Y. Speech recognition in noisy environments: A survey. *Speech Commun.* **1995**, *16*, 261–291. [\[CrossRef\]](#)
18. Juang, B. Speech recognition in adverse environments. *Comput. Speech Lang.* **1991**, *5*, 275–294. [\[CrossRef\]](#)
19. Zhang, Z.; Geiger, J.; Pohjalainen, J.; Jin, W.; Schuller, B. Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Trans. Intell. Syst. Technol.* **2018**, *9*, 1–28. [\[CrossRef\]](#)
20. Li, J.; Deng, L.; Gong, Y.; Haeb-Umbach, R. An Overview of Noise-Robust Automatic Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Proc.* **2014**, *22*, 745–777. [\[CrossRef\]](#)
21. Upadhyay, N.; Rosales, H.G. Robust Recognition of English Speech in Noisy Environments Using Frequency Warped Signal Processing. *Natl. Acad. Sci. Lett.* **2018**, *41*, 15–22. [\[CrossRef\]](#)
22. Kang, B.O.; Jeon, H.B.; Park, J.G. Speech Recognition for Task Domains with Sparse Matched Training Data. *Appl. Sci.* **2020**, *10*, 6155. [\[CrossRef\]](#)
23. Zheng, F.; Zhang, G.; Song, Z. Comparison of different implementations of MFCC. *J. Comput. Sci. Technol.* **2001**, *16*, 582–589. [\[CrossRef\]](#)
24. Nassif, A.B.; Shahin, I.; Attili, I.; Azzeh, M.; Shaalan, K. Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access* **2019**, *7*, 19143–19165. [\[CrossRef\]](#)
25. Raj, B.; Stern, R. Missing-feature approaches in speech recognition. *IEEE Signal Process. Mag.* **2005**, *22*, 101–116. [\[CrossRef\]](#)
26. Gupta, K.; Gupta, D. An analysis on LPC, RASTA and MFCC techniques in Automatic Speech recognition system. In Proceedings of the 2016 6th International Conference—Cloud System and Big Data Engineering (Confluence), Noida, India, 14–15 January 2016; pp. 493–497.
27. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462.
28. Anusuya, M.A.; Katti, S.K. Front end analysis of speech recognition: A review. *Int. J. Speech Technol.* **2011**, *14*, 99–145. [\[CrossRef\]](#)
29. Lee, K.H.; Kang, W.H.; Kang, T.G.; Kim, N.S. Integrated DNN-based model adaptation technique for noise-robust speech recognition. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5245–5249. [\[CrossRef\]](#)
30. Ko, T.; Peddinti, V.; Povey, D.; Khudanpur, S. Audio augmentation for speech recognition. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association—Interspeech 2015, Dresden, Germany, 6–10 September 2015.
31. Nguyen, T.-S.; Stuker, S.; Niehues, J.; Waibel, A. Improving Sequence-To-Sequence Speech Recognition Training with On-The-Fly Data Augmentation. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7689–7693. [\[CrossRef\]](#)
32. Prisyach, T.; Mendelev, V.; Ubskiy, D. Data Augmentation for Training of Noise Robust Acoustic Models. In *International Conference on Analysis of Images, Social Networks and Texts*; Springer: Cham, Switzerland, 2016; pp. 17–25. [\[CrossRef\]](#)
33. Shahnawazuddin, S.; Adiga, N.; Kathania, H.K.; Sai, B.T. Creating speaker independent asr system through prosody modification based data augmentation. *Pattern Recognit. Lett.* **2020**, *131*, 213–218. [\[CrossRef\]](#)
34. Staš, J.; Hladek, D.; Pleva, M.; Juhar, J. Slovak language model from Internet text data. In *Towards Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues, LNCS 6456*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 352–358.

35. Byrne, W.; Hajič, J.; Ircing, P.; Jelinek, F.; Khudanpur, S.; McDonough, J.; Peterek, N.; Psutka, J. Large Vocabulary Speech Recognition for Read and Broadcast Czech. In Proceedings of the Text, Speech and Dialogue—Second International Workshop, TSD'99, Plzen, Czech Republic, 13–17 September 1999; pp. 235–240. [\[CrossRef\]](#)
36. Ircing, P.; Krbec, P.; Hajic, J.; Psutka, J.; Khudanpur, S.; Jelinek, F.; Byrne, W. On large vocabulary continuous speech recognition of highly inflectional language-Czech. In Proceedings of the Seventh European Conference on Speech Communication and Technology, Aalborg, Denmark, 3–7 September 2001.
37. Maucec, M.S.; Kacic, Z.; Horvat, B. A framework for language model adaptation for highly-inflected Slovenian language. In *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*; ISCA: Sophia Antipolis, France, 2001.
38. Schwenk, H. Trends and challenges in language modeling for speech recognition and machine translation. In Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition & Understanding, Moreno, Italy, 13 November–17 December 2009; p. 23. [\[CrossRef\]](#)
39. Mousa, A.E.-D.; Shaik, M.A.B.; Schlüter, R.; Ney, H. Morpheme level hierarchical pitman-yor class-based language models for LVCSR of morphologically rich languages. In Proceedings of the Annual Conference of the International Speech Communication Association, Lyon, France, 25–29 August 2013. [\[CrossRef\]](#)
40. Staš, J.; Hladek, D.; Juhar, J. Morphologically motivated language modeling for Slovak continuous speech recognition. *J. Electr. Electron. Eng.* **2012**, *5*, 233–237.
41. Donaj, G.; Kačič, Z. Context-dependent factored language models. *EURASIP J. Audio, Speech, Music Process.* **2017**, *2017*, 6. [\[CrossRef\]](#)
42. Vazhenina, D.; Markov, K. Factored language modeling for Russian LVCSR. In Proceedings of the International Joint Conference on Awareness Science and Technology and Ubi-Media Computing, iCAST 2013 and UMEDIA 2013, Aizu-Wakamatsu, Japan, 2–4 November 2013.
43. Maucec, M.S.; Rotovnik, T.; Zemljak, M. Modelling Highly Inflected Slovenian Language. *Int. J. Speech Technol.* **2003**, *6*, 245–257. [\[CrossRef\]](#)
44. Karpov, A.; Kipyatkova, I.; Ronzhin, A. Very large vocabulary ASR for spoken Russian with syntactic and morphemic analysis. In Proceedings of the Annual Conference of the International Speech Communication Association, Florence, Italy, 27–31 August 2011. [\[CrossRef\]](#)
45. Pipiras, L.; Maskeliūnas, R.; Damaševičius, R. Lithuanian Speech Recognition Using Purely Phonetic Deep Learning. *Computers* **2019**, *8*, 76. [\[CrossRef\]](#)
46. Polat, H.; Oyucu, S. Building a Speech and Text Corpus of Turkish: Large Corpus Collection with Initial Speech Recognition Results. *Symmetry* **2020**, *12*, 290. [\[CrossRef\]](#)
47. Rotovnik, T.; Maučec, M.S.; Kačič, Z. Large vocabulary continuous speech recognition of an inflected language using stems and endings. *Speech Commun.* **2007**, *49*, 437–452. [\[CrossRef\]](#)
48. Zgank, A.; Verdonik, D.; Markus, A.Z.; Kacic, Z. BNSI Slovenian broadcast news database—Speech and text corpus. In Proceedings of the Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005. [\[CrossRef\]](#)
49. Vicsi, K.; Szaszák, G. Using prosody to improve automatic speech recognition. *Speech Commun.* **2010**, *52*, 413–426. [\[CrossRef\]](#)
50. Bang, J.-U.; Kim, S.-H.; Kwon, O.-W. Acoustic Data-Driven Subword Units Obtained through Segment Embedding and Clustering for Spontaneous Speech Recognition. *Appl. Sci.* **2020**, *10*, 2079. [\[CrossRef\]](#)
51. Theera-Umpon, N.; Chansareewittaya, S.; Auephanwiriyakul, S. Phoneme and tonal accent recognition for Thai speech. *Expert Syst. Appl.* **2011**, *38*, 13254–13259. [\[CrossRef\]](#)
52. Verdonik, D. Between understanding and misunderstanding. *J. Pragmat.* **2010**, *42*, 1364–1379. [\[CrossRef\]](#)
53. Lopes, C.; Perdigao, F. Broad phonetic class definition driven by phone confusions. *EURASIP J. Adv. Signal Process.* **2012**, *2012*, 158. [\[CrossRef\]](#)
54. Golik, P.; Tüske, Z.; Schlüter, R.; Ney, H. Development of the RWTH transcription system for slovenian. In Proceedings of the 14th Annual Conference of the International Speech Communication Association, Lyon, France, 25–29 August 2013. [\[CrossRef\]](#)
55. Pleva, M.; Čížmar, A.; Juhar, J.; Ondaš, S.; Mirilovič, M. Towards Slovak Broadcast News Automatic Recording and Transcribing Service. In *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction, Lecture Notes in Computer Science 5042*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 158–168.
56. Prochazka, V.; Pollak, P.; Zdansky, J.; Nouza, J. Performance of Czech Speech Recognition with Language Models Created from Public Resources. *Radio Eng.* **2011**, *20*, 1002–1008.
57. Vizslay, P.; Staš, J.; Kočtúr, T.; Lojka, M.; Juhár, J. An extension of the Slovak broadcast news corpus based on semi-automatic annotation. In Proceedings of the Tenth International Conference on Language Resources and Evaluation—LREC 2016, Portorož, Slovenia, 23–28 May 2016; pp. 4684–4687.
58. Nouza, J.; Safarik, R.; Cerva, P. ASR for South Slavic Languages Developed in Almost Automated Way. In Proceedings of the 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016. [\[CrossRef\]](#)