Designing the Latvian Speech Recognition Corpus

Mārcis Pinnis[1], Ilze Auziņa[2], Kārlis Goba[1]
[1]Tilde, Vienibas gatve 75a, Riga, Latvia
[2]Institute of Mathematics and Computer Science, University of Latvia, 29 Raina Blvd., Riga, Latvia
E-mail: marcis.pinnis@tilde.lv, ilze.auzina@lumii.lv, karlis.goba@tilde.lv

Abstract

In this paper the authors present the first Latvian speech corpus designed specifically for speech recognition purposes. The paper outlines the decisions made in the corpus designing process through analysis of related work on speech corpora creation for different languages. The authors provide also guidelines that were used for the creation of the Latvian speech recognition corpus. The corpus creation guidelines are fairly general for them to be re-used by other researchers when working on different language speech recognition corpora. The corpus consists of two parts – an orthographically annotated corpus containing 100 hours of orthographically transcribed audio data and a phonetically annotated corpus containing 4 hours of phonetically transcribed audio data. Metadata files in XML format provide additional details about the speakers, noise levels, speech styles, etc. The speech recognition corpus is phonetically balanced and phonetically rich and the paper describes also the methodology how the phonetical balancedness has been assessed.

Keywords: speech recognition corpus, Latvian, corpus creation guidelines

1. Introduction

An annotated speech corpus is a necessary prerequisite for both speech recognition and speech synthesis research and system development. Without an annotated speech corpus the possibilities to design and implement systems for a specific language are limited. However, as shown by Schultz and Waibel (1997), this limitation can be overcome by applying acoustic model bootstrapping techniques. In recent years this technique has been actively applied by researchers of the Quaero project (Lamel, 2012; Adda-Decker et al., 2010), which aims at implementing broadcast news speech recognition systems for all official languages of the European Union. They use carefully selected seed acoustic models of different languages and adapt them with the help of target language speech and text corpora. Although the bootstrapped systems can achieve a word error rate (WER) of less than 20%, the systems are focussed to a narrow speaker base (broadcast news reporters). The speaker base of every language differs enormously because of various accentual, dialectal, physical, and many other characteristics. Therefore, the authors' goal was to create an orthographically (up to 100 hours) and phonetically (up to four hours) annotated speech corpus (the Latvian Speech Recognition Corpus) that would represent the major speaker base of Latvian and would allow widening research activities in speech recognition and subsequently also speech synthesis of Latvian.

Creation of a speech recognition corpus that represents the major speaker base of a language is a very challenging task, therefore, the authors have structured the paper so that it can be used as guidelines for different language corpora creation, an aspect that in related works on speech corpora has not been sufficiently addressed. The authors are, of course, not the first ones to create a national speech corpus for a specific language. To name, but a few, related works on speech corpora creation are Oostdijk (2000) on Dutch, Johannessen et al. (2007) on Norwegian, Stănescu et al. (2012) on Romanian, and many, many others.

Latvian belongs to the Baltic language group, it is one of the official languages of the European Union, the state language of Latvia and it is daily used by approximately 2 million people worldwide. Latvian is a language with rich morphology and a relatively free word order. Because of the lack of an annotated speech corpus, research in speech recognition on Latvian has been very limited. To the best of our knowledge the only publicly known speech recognition efforts have been carried out by Oparin et al. (2013) on broadcast news speech recognition using acoustic model bootstrapping methods. Speech synthesis, on the other hand, has received more attention – two of the better known systems have been created by Goba & Vasiļjevs (2007), and Pinnis & Auziņa (2012).

The paper is further structured as follows: section 2 describes different criteria (the acoustic quality of speech, phonetic coverage, linguistic criteria, and speaker variability) that the authors have set for the speech corpus, section 3 gives a brief introduction in the speech corpus annotation guidelines. The paper is concluded in section 4.

2. Criteria for the Design of the Speech Corpus

The Latvian Speech Recognition Corpus has been designed to satisfy a set of criteria, which specify the required quality of speech data and the proportional distribution of data with different speaker characteristics. In this paper the authors provide a brief overview of the criteria defined for the speech corpus (extended reasoning behind each criteria will be provided in the full paper). The criteria are as follows:

- ***Audio signal quality***. Similarly to the general trend of broadband speech recognition system development

(Amdal et al., 2008; Federico et al., 2000; Oostdijk, 2000; and many others), the speech audio data have to have a minimum frequency of 16 kHz with 16 bits allocated per sample. Earlier research by Barras et al. (2001) has shown that audio signal compression (including different types of compression) may significantly lower speech recognition quality. This means that if compressed audio data are used in the creation of the corpus, in a decompressed (and down-sampled) state it has to be equivalent to the minimum quality requirements. Furthermore, for our corpus we deliberately did not consider telephone quality (e.g., 8kHz and 8 bit per sample) speech because of two main reasons: 1) earlier research by Weintraub et al. (1994) has shown that higher quality speech audio data can be effectively down-sampled to lower quality speech audio data without a significant loss of speech recognition quality, and 2) current mobile phones (at least the vast majority) already support broadband quality speech recording and telephone quality audio data are becoming obsolete. Finally, we did not set restrictions on the recording hardware. As shown by Suominen et al., 2013, different speech recording devises, including smartphones, media players, tablet computers, etc., can have a significant effect on the speech recognition quality (in terms of word error rate) and the quality differs notably between different devices. However, multi-condition training (Rajnoha, 2009) could potentially allow limiting the negative effects caused by the recording devices and, thus, we do not limit the audio recordings to some specific microphones or recording devices.

- ***Distribution of noise***. In order to be able to develop speech recognition systems that can be executed in different environments and that can be used for wide purposes, we should be able to build speech acoustic models that are robust towards the noise representative to the different environments. Recent research in speech recognition (Gemmeke et al., 2011; Rajnoha, 2009) has shown that multi-condition training achieves higher speech recognition performance in various environmental conditions. As shown by Stouten (2006), Yapanel et al. (2001) and others, usage of just clean speech audio data for acoustic model training will cause the speech recognition quality to drop quite significantly in noisy environments. Therefore, the Latvian speech recognition corpus has been developed to include speech data with different types of background noise (office, street, in-car, etc. noise) with different signal-to-noise ratios (SNR). The majority of the data, however, contain a relatively low level of noise (the SNR being between 15-25dB). The background noise types for each utterance are identified in the corpus metadata that describes each audio recording. Overlapping speech segments (with two or more simultaneous speakers) are not included in the corpus.
- ***Phonetic coverage***. Following recent research on other language speech recognition corpora (Abushariah et al., 2012; Irtza & Hussain, 2013; Stănescu et al., 2012) the corpus has been designed to be phonetically balanced in order to be representative of natural speech and phonetically rich so that the trained acoustic models (up to triphone models) would efficiently generalize over different speakers with different characteristics. The analysis of the created corpora with respect to the phonetic balancedness is described in section 4.
- ***Speech styles***. The corpus consists of prepared speech (40%) and spontaneous speech (60%). Prepared speech covers TV and radio news, audiobooks, publicly read speeches, read presentation, etc. The spontaneous speech covers TV and radio discussions, interviews, recorded conversations, speeches according to a prepared plan (but not read speeches), e.g., presentation, lecture speeches, etc. The corpus covers all the different speech styles because of two main reasons: 1) to have a larger speech data diversity in terms of coverage of filler words, differences in the speed of the pronounced words, larger diversity of intonations, etc., and 2) to later be able to tune speech recognition engines for specific speech recognition tasks (e.g., dictation transcription, broadcast news transcription, lecture transcription, etc.).
- ***Physical characteristics of speakers***. The speech corpus is representative of speakers of both genders in equal proportions and contains speech from speakers of different ages. Following research by Johannessen et al. (2007), the speakers are grouped in three age groups: from 16 to 25 (up to 25%), from 26 to 50 (at least 50%), and from 51 to 75 (up to 25%). The upper boundary is lower than that of Johannessen et al. (2007) as our corpus is relatively small and we did not have enough speakers to cover that age group. However, earlier research does not necessarily give a clear insight of why such division has been used. E.g., the works of Guðnason et al. (2012), Oostdijk (2000), Pineda et al. (2009), and Sarfraz et al. (2010) use different speaker age intervals, however, they do not justify them. Our reasoning behind following the division of Johannessen et al. (2007) with the limited upper boundary is as follows: we excluded the speech of children and people older than 75 years because the speech characteristics in these groups are quite different and require training of separate acoustic models. The remaining groups are divided in: 1) young people, 2) people in the active working ages (the potential user base of speech recognition technologies), 3) older people who could be potential users of speech recognition technologies in the future.
- ***Linguistic characteristics of speakers***. The Latvian Speech Recognition Corpus consists of utterances from the Latvian literary (formal) language, however, pronounced by a variety of speakers, including speakers with different dialectal (up to 15%; e.g., speakers of the Livonian dialect or the High Latvian (*Latgalian*) dialect) or accentual (up to 25%; e.g.,

Belorussian, English, Russian, Ukrainian) characteristics. At least 60% of the data are from speakers without any dialectal or accentual characteristics.

- ***Data formats***. The corpus consists of speech audio files, multiple meta-data XML documents, and label files for the phonetically annotated corpus. The audio files are 1 channel, 16 bit WAV files. The frequency varies depending on the source audio data quality (with a minimum of 16 KHz). There are in total three meta-data XML documents: 1) meta-data of classifiers (i.e., the possible background noise types, speech styles, age groups, and accent types) that are used in the other two XML documents; 2) speaker meta-data, which (anonymously) lists all speakers in the speech recognition corpus with their age group, gender, and accent group; 3) orthographic annotation meta-data, which contains all orthographic transcriptions of all speech audio files. Figure 1 shows that the orthographic annotation metadata XML document consists of entries for audio recordings (which can be full interviews, news broadcasts, lectures, etc.). The recordings are segmented into fragments. Each fragment contains speech of only one speaker. Each fragment is divided in parts, which consist of either speech utterances or longer filler segments (including silence segments).
- ***Corpus size***. The total size of the corpus is 100 hours of orthographically annotated speech and 4 hours of phonetically annotated speech. The statistics are given in Section 4.

```
<?xml version="1.0"
encoding="UTF-8"?>
<files>
 <file name="[Recording Name]">
  <fragment length="9.01" speaker="1"
type="4" place="1" snr="20.46">
   <part length="2.25"
audio file="audio/[Recording Name]/p
art 1.wav" >bet mēs turpinām ar citām
aktualitātēm</part>
   <part length="0.70"
audio_file="audio/[Recording Name]/p
art 2.wav" >(.hh)</part>
  </fragment>
 </file>
</files>
```

Figure 1**:** An example of the orthographic annotation meta-data

3. Speech Corpus Annotation Guidelines

This section briefly describes the speech corpus annotation guidelines that were used by the annotators of the Latvian Speech Recognition Corpus. The corpus contains speech annotation in two levels: orthographic annotation and phonetic annotation. The next two subsections describe both annotation levels.

3.1. Orthographic Annotation Guidelines

Following earlier research on orthographically annotated speech corpora creation (Goedertier et al., 2000; Johannessen et al., 2007; Oostdijk et al., 2002) the authors have created a set of rules for the orthographic transcription. The rules specify how to annotate (also expand if necessary) numerals, abbreviations, punctuations, non-speech fragments (e.g., breathing, filled stops, laughter, etc.), abrupt words, unclear speech, words spoken in a different language (e.g., named entities, quotes, etc.), non-verbal elements (e.g., singing, speech while inhaling, exhaling, or laughing, etc.), physiological noise (e.g., snuffling, smacking, coughing, etc.), background noise and other types of information characterising a speech fragment within an utterance. Several of the acoustic event categories are listed in Table 1 and an example of an orthographically annotated utterance is given in Figure 2.

	Type	Label
Speaker noise	Inhalation	(.h)
	Exhalation	(h.)
	Vocal hesitation	(ē), (ā), (em)
	Whisper	<čuksts> text </čuksts>
	Laugh	@
		<@> text </@>
Pauses	Short pause (> 300 ms)	(.)
	Long pause (> 1 sec)	For example,(0.5)
Transient background noise	Physiological noise	<ftr/>
		<ftr> text </ftr>
	Music	<muz> text </muz>
	Other	<tr/>
		<tr> text </tr>

Table 1: Categories of acoustic events in the orthographic annotation

```
<part length="4.397" audio file=
"audio/TV_Show_1.wav" >(ē) valodu viņa
izjuta arī kā brīnumu, katrs vārds
viņai bij [bija] tā kā</part>
```

Figure 2: An example of an orthographically annotated utterance

The example above depicts an utterance of 4.397 seconds starting with a filled stop ("(ē)" – phonetically /æ:/ or /e:/), followed by grammatically correctly pronounced words, and one word pronounced differently from the grammatically correct form (the correct form is given in square brackets). We followed the natural segmentation (in other words, the speech division into intonational phrases at full stops, inhalations/exhalations, etc.) of the speech in order to segment speech recordings in utterances. Each utterance corresponds to an individual audio file. In the orthographic transcription we use the pronunciation of words as the main spelling (regardless of the grammatical correctness), but in cases if there is a deviation from the norms of the standard orthography of

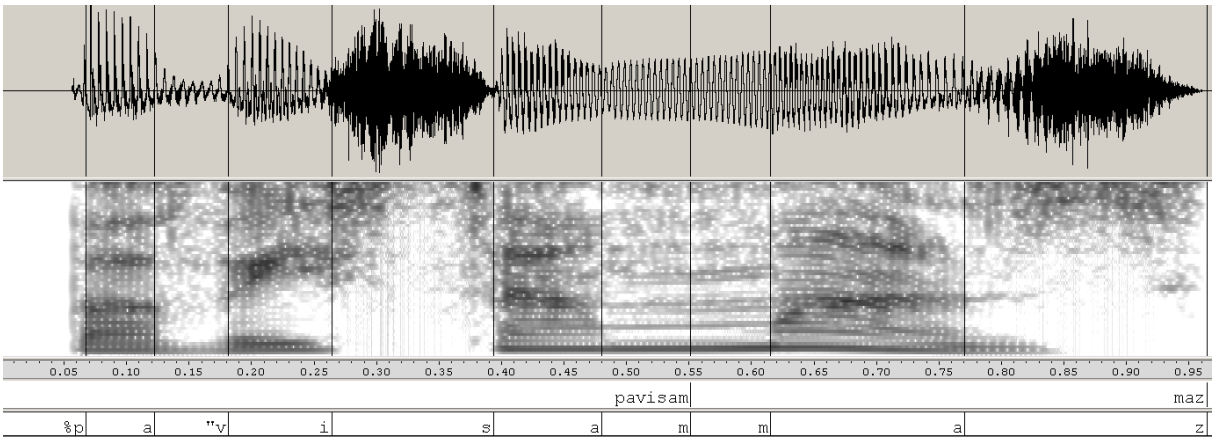

Figure 3: An example of a phonologically annotated utterance *pavisam maz* 'very little'

the Latvian language, the correct form is given in square brackets, for example,

1. Numerals are shortened, *trīesnt*, *trīesmit*, *trīsmit*, *trīsdesmit*, *trīsnt* [*trīsdesmit*] 'thirty', *čēesmit* [*četrdesmit*] 'forty', *sēešmit* [*sešdesmit*] 'sixty', *pieesmit* [*piecdesmit*] 'fifty' (as can be seen from the examples digits have alternate pronunciation forms and the orthographic transcription accurately reflects the form that is actually pronounced);
2. Adverbs *tad* 'then', *kad* 'when' are pronounced without consonants *t*, *d*: *ta* [*tad*] *ka* [*kad*];
3. Person forms of verbs *būt*, *nebūt*, *vajadzēt*, *nevajadzēt* are shortened: *vaig* [*vajag*] 'need', *nevaig* [*nevajag*] 'do not need', *bij* [*bija*] 'was/were', *nebij* [*nebija*] 'was not/were not', *esu* [*esmu*] '(I) am';
4. Some international words are pronounced with a long vowel instead of a short one (*rādio* [*radio*] 'radio') or the opposite (*muzika* [*mūzika*] 'music'); vowels can also be reduced (*intresants* [*interesants*] 'interesting').

Capital letters are used in proper names (e.g., *Rēzeknes Universitāte* 'University of Rezekne') and acronyms (e.g., *LETA*, *NATO*) only. Abbreviations are represented by their full orthographic forms, unless they are spoken in their abbreviated form, e.g., *LETA*, *ANO* 'UN – United Nations'. If a speaker pronounces letters, acronyms, internet addresses or abbreviations in words, traditional writing is given into the brackets, e.g. *el vē [2, LV]*; *vē vē vē punkts rēzekne punkts el vē [8, www.rezekne.lv]*. Numbers and symbols are written out as words (e.g., *septini procenti* '7%'). Text codes (diacritic marks) are used to mark mispronunciations (e.g., *mēs būš- būtu atra- atražojuši tikai paši*) and unclear text (e.g., *nu līdz šim mēs esam dzirdējušas {--} labas atsauksmes*). Punctuation marks are restricted to comma, period, and question mark only.

3.2. Phonetic Annotation Guidelines

The phonetically Annotated Latvian Speech Corpus consists of four hours of speech from 67 speakers (36 female and 31 male). The corpus is provided in the broad transcription (or phonemic transcription): transcription that relates the allophones produced by the speakers to the phonemes of Latvian. However additional information about phonetic variations of some specific allophones in utterances is also marked, e.g., we use different symbols for:

- phoneme /ɑ/ in *masa* 'mass' (the second /ɑ/ becomes extra short);
- phoneme /s/ in Sg. Nom. *kase* 'booking-office' and Sg. Loc. *kasē* 'in the booking-office' (the first word has a twice longer /s/), etc.

The set of symbols used in the machine-readable phonetic transcriptions has been derived from the SAMPA [1] set (Gibbon, et al., 1997).

The annotation by human annotators is performed in a partially automatic fashion. At first, the phonemic transcription is generated automatically from the orthographic transcription using context sensitive grapheme-to-phoneme rules. Then the phonemic transcriptions are automatically aligned with the speech signal. Finally, the transcribed data are manually verified by human annotators both on the word and the phonemic level. Figure 3 shows an example of the final annotation process performed using WaveSurfer (Sjölander & Beskow, 2000).

4. Phonetic Balancedness of the Speech Recognition Corpus

In order to verify that the speech recognition corpus is phonetically balanced, we compared the relative distribution of triphones in the orthographically annotated corpus with a large reference corpus of 120 million running words. The reference corpus consists of news articles, fiction, legislative documents, etc. We used triphones, because 1) they are the minimal phonetic segments that are required to capture the acoustic changes of a phoneme between two adjacent phonemes, and 2) they are often used in speech recognition as the leading acoustic segments.

Because the orthographically annotated corpus as well as the reference corpus do not contain phonetic transcriptions, we used the rule-based phonetic transcription tool from the Tilde's text-to-speech system Visvaris (Goba and Vasiļjevs, 2007) in order to acquire an approximation of the phonetic coverage in both corpora. The phonetic sequences were then converted into triphone

[1] The SAMPA computer readable phonetic alphabet is available online at:
http://www.phon.ucl.ac.uk/home/sampa/home.htm

sequences (not taking into account spaces between words, which corresponds to natural speech). The triphone statistics of both corpora are given in Table 2.

	Audio corpus	Reference corpus
Triphones in total	5.8 million	578 million
Unique triphones	8 302	14 851

Table 2: The triphone statistics in the orthographically annotated corpus and the reference text corpus

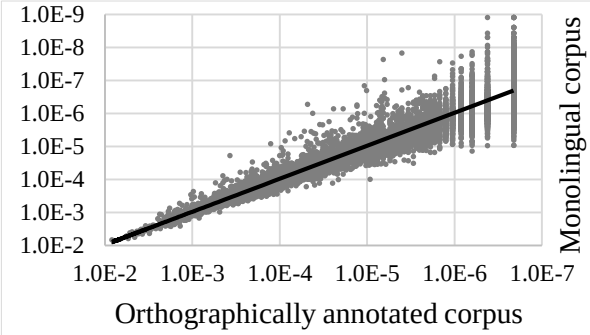

Figure 4: Triphone relative distribution comparison between the orthographically annotated corpus and the reference text corpus

The triphone relative frequency distributions in both corpora were further compared. The Figure 4 visually depicts how similar the distributions are. In order to verify how much the triphone distribution between the two corpora correlates, we calculated also the Pearson product-moment correlation coefficient. The positive result of 0.9753 indicates of a relatively high correlation, thereby we conclude that the orthographically annotated speech recognition corpus is phonetically balanced.

5. Speech Corpus Statistics

The development of the Latvian Speech Recognition Corpus has been finalised at the end of 2013. The statistics of the orthographically annotated data in the Latvian Speech Recognition Corpus are given in Table 3; the proportional distribution of speech data with respect to speaker gender and age is given in Figure 5 and Figure 6. Figure 7 depicts the proportional distribution of the data with respect to the style of speech. The total length of the Latvian Speech Recognition Corpus is 100 hours and 1 minute. It includes both verbal and non-verbal segments (see Table 4). Most audio data included in the corpus have a frequency of 44.1 kHz with 16 bits allocated per sample.

Number of unique words	~72.5 k
Number of running words	~837 k
Total number of speakers	1 851
Men	1 016
Women	835

Table 3: The statistics of the Latvian Speech Recognition Corpus

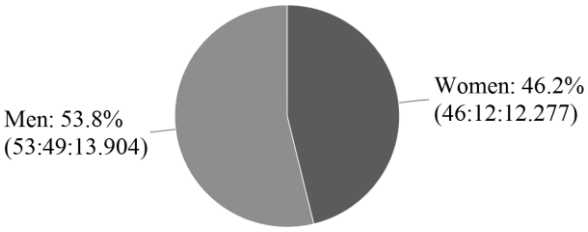

Figure 5: Data distribution with respect to the gender of speakers

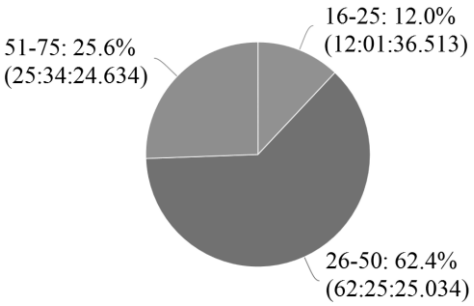

Figure 6: Data distribution with respect to the age of speakers

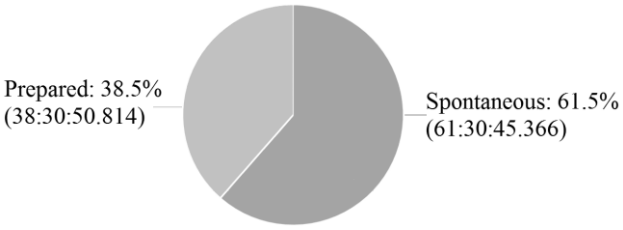

Figure 7: Data distribution with respect to speech styles

Type	Total length
Inhalation, exhalation	3 h 45 min (13 538 s)
Pauses	1 h 55 min (6 911 s)
Non-verbal segments	19 min (1 137 s)
Verbal segments	94 h 1 min (338 500 s)
The whole corpus	100 h 1 min (360 086 s)

Table 4: Data distribution with respect to different speech segment types

The data included in the corpus have been selected by ensuring that the audio data contain different noise types: (1) audio data without background noise (inside a studio / outside a studio without background noise), (2) data recorded in a studio, but with background noise (e.g., physiological noise), (3) outside a studio with background noise (e.g., office noise), (4) street noise (e.g., noise caused by vehicles, pedestrians, etc.), (5) noise inside a car, and (6) loud music as background noise. The distribution of different levels of noise in the orthographically annotated speech recognition corpus in terms of signal-to-noise ratio estimated on whole speech

segments (containing multiple utterances) is given in Figure 8.

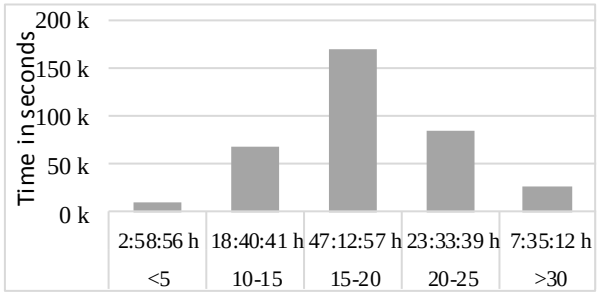

Figure 8: Distribution of the orthographically annotated corpus with respect to different SNR levels (in dB)

6. Conclusion and Future Work

In this paper the authors have presented the overall design of the Latvian Speech Recognition Corpus. The corpus consists of 100 hours of orthographically annotated speech audio data and 4 hours of phonetically annotated speech audio data. The corpus is both phonetically rich and balanced. The paper also described the reasoning behind different choices made during the development of the corpus as well as described the methodology that was used in order to verify the phonetic balancedness of the speech recognition corpus. Therefore, the authors believe that the paper can serve as guidelines for other researchers who develop or want to develop speech recognition corpora for under-resourced languages.

The Latvian Speech Recognition Corpus in the years to come will be an asset for further research in speech recognition (and also speech synthesis) of Latvian – a language that did not have its dedicated speech recognition corpus before.

In addition the speech corpus will play a crucial role in linguistic research, for example, comparing pronunciation by men and women; flapping across word boundaries in spontaneous speech, the omission of sounds (sound deletion), assimilation across word boundaries, etc.

Further statistics of the Latvian Speech Recognition Corpus can be found on the project's home page http://runa.korpuss.lv.

7. Acknowledgements

The research leading to these results has received funding from the research project "Information and Communication Technology Competence Center" of EU Structural funds, contract nr. L-KC-11-0003 signed between ICT Competence Centre and Investment and Development Agency of Latvia, Research No. 2.9 "Speech corpus creation, principles, methods, realisation".

8. References

Abushariah, M., Ainon, R., Zainuddin, R., Elshafei, M., & Khalifa, O. (2012). Arabic Speaker-Independent Continuous Automatic Speech Recognition Based on a Phonetically Rich and Balanced Speech Corpus. *The International Arab Journal of Information Technology*, 9(1), 84--93.

Adda-decker, M., Lamel, L., & Snoeren, N. D. (2010). Initializaing Acoustic Phone Models of Under-resourced Languages: A Case-study of Luxembourgish. *Proceedings of the second International Workshop on Spoken Languages Technologies for Under-resourced languages* (*SLTU10*) (pp. 74--80). Penang, Malaysia.

Amdal, I., Strand, O. M., Almberg, J., & Svendsen, T. (2008). RUNDKAST : An Annotated Norwegian Broadcast News Speech Corpus. *Proceedings of LREC 2008*.

Barras, C., Lamel, L., & Gauvain, J.-L. (2001). Automatic Transcription of Compressed Broadcast Audio. *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)* (Vol. 1, pp. 265–268). Salt Lake City, UT, USA: Ieee. doi:10.1109/ICASSP.2001.940818.

Federico, M., Giordani, D., & Coletti, P. (2000). Development and evaluation of an Italian Broadcast News Corpus. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*.

Gemmeke, J. F., Segbroeck, M. Van, Wang, Y., Cranen, B., & Van Hamme, H. (2011). Automatic Speech Recognition Using Missing Data Techniques: Handling of Real-World Data. In D. Kolossa & R. Häb-Umbach (Eds.), *Robust Speech Recognition of Uncertain or Missing Data: Theory and Applications* (pp. 157--185). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-21317-5_7.

Gibbon, D., Moore, R., & Winski, R. (1997). SAMPA computer readable phonetic alphabet, Part IV, section B. *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, Berlin and New York.

Goba, K., and Vasiljevs, A. (2007). Development of Text-To-Speech System for Latvian. In Joakim Nivre, H.-J. Kaalep, K. Muischnek, & M.Koit (Eds.), *In Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007* (pp. 67–-72). Tartu, Estonia.

Goedertier, W., Goddijn, S., & Martens, J. (2000). Orthographic Transcription of the Spoken Dutch Corpus. *Proceedings of LREC-2000* (pp. 909--914).

Guðnason, J., Kjartansson, O., Jóhannsson, J., Carstensdóttir, E., Vilhjálmsson, H. H., Loftsson, H., Helgadóttir, S., et al. (2012). ALMANNARÓMUR: An Open Icelandic Seech Corpus. In *Spoken Language Technologies for Under-Resourced Languages*.

Irtza, S., & Hussain, S. (2013). Minimally balanced corpus for speech recognition. *Proceedings of the 1st International Conference on Communications, Signal*

Processing, and their Applications (*ICCSPA'13*). Sharjah, UAE.

Johannessen, J.B., Hagen, K., Priestley, J.J., and Nygaard, L. (2007). An Advanced Speech Corpus for Norwegian. *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007* (pp. 29--36). Tartu, Estonia. ISBN 978-9985-4-0513-0.

Lamel, L. (2012). Multilingual Speech Processing Activities in Quaero: Application to Multimedia Search in Unstructured Data. *Proceedings of the Fifth International Conference: Human Language Technologies-The Baltic Perspective*.

Oostdijk, N. (2000). The Spoken Dutch Corpus. Overview and first Evaluation. *Proceedings of LREC 2000* (pp. 887--894). Athens, Greece.

Oostdijk, N., Goedertier, W., Eynde, F. Van, Boves, L., Martens, J., Moortgat, M., & Baayen, H. (2002). Experiences from the Spoken Dutch Corpus Project. *Proceedings of LREC 2002* (pp. 340--347).

Oparin, I., Lamel, L., & Gauvain, J. (2013). Rapid Development of a Latvian Speech-to-text System. *Proceedings of ICASSP'13* (pp. 2--6). Vancouver, Canada.

Pinnis, M., & Auziņa, I. (2010). Latvian Text-to-Speech Synthesizer. In I. Skadiņa & A. Vasiļjevs (Eds.), *Proceedings of the 2010 conference on Human Language Technologies--The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010* (pp. 69--72). Riga, Latvia: IOS Press. doi:10.3233/978-1-60750-641-6-69

Rajnoha, J. (2009). Multi-Condition Training for Unknown Environment Adaptation in Robust ASR Under Real Conditions. *Acta Polytechnica*, 49(2-3), 3–7.

Sarfraz, H., Hussain, S., Bokhari, R., Raza, A. A., Ullah, I., Pervez, S., Mustafa, A., et al. (2010). Speech Corpus Development for a Speaker Independent Spontaneous Urdu Speech Recognition System. *Proceedings of O-COCOSDA 2010*. Kathmandu, Nepal.

Schultz, T., & Waibel, A. (1997, September). Fast bootstrapping of LVCSR systems with multilingual phoneme sets. *In the Fifth European Conference on Speech Communication and Technology- Eurospeech 1997*.

Sjölander, K., & Beskow, J. (2000, October). Wavesurfer-an open source speech tool. *In INTERSPEECH* (pp. 464--467).

Stănescu, M., Cucu, H., Buzo, A., & Burileanu, C. (2012). ASR for Low-resourced Languages: Building a Phonetically Balanced Romanian Speech Corpus. *Proceedings of the 20th European Signal Processing Conference* (*EUSIPCO 2012*) (pp. 2060--2064). Bucharest, Romania.

Stouten, V. (2006). *Robust automatic speech recognition in time-varying environments*. *KU Leuven, Diss*. KU Leuven.

Weintraub, M., Neumeyer, L., & Park, M. (1994). Constructing telephone acoustic models from a high-quality speech corpus. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-94)* (Vol. i, p. I/85–I/88 vol.1). doi:10.1109/ICASSP.1994.389349.

Yapanel, U., Hansen, J. H. L., Sarikaya, R., & Pellom, B. (2001). Robust Digit Recognition in Noise : An Evaluation Using the AURORA Corpus. *Proceedings of EUROSPEECH 2001* (pp. 209–212).