The Potential of the Lithuanian Speech Corpus

Term Paper, GSLT-1 level Speech Technology course, autumn 2005

Dainora Kuliešienė, Gintarė Grigonytė

Vytautas Magnus University, Lithuania

We present Vytautas Magnus University (VMU) *Speech Corpus* of the Lithuanian language in this paper. The aim of creating the corpus was enabling Lithuanian spoken language researchers to use popular tools like HTK, MBROLA etc. VMU Speech Corpus is built upon 4 speakers (2 males and 2 females) voice records, each record has a the same set of around 740 isolated words. The corpus includes *time*-aligned phone-level and *word-level* transcriptions data. The vocabulary of the corpus has been carefully chosen and includes all distinct and independent Lithuanian sounds (275 phonetic units total) such as phonemes and phoneme clusters (phonetic units), at this point VMU Speech Corpus of the Lithuanian language becomes applicable for a big number of researches. We will shortly describe problems related to the file structure of the corpus and ASCII coding of Lithuanian annotations, corpus validation and standardization.

Introduction

A very rough categorization of the speech corpuses might distinct *specialized* and *universal* corpuses. *Specialized* corpuses are oriented to particular features of speech: dialects; male/female/children; spontaneous speech; reading of the text; individual words; commands and continuous speech. While the *universal* corpuses cover many common features of speech.

We will present *universal* annotated Speech Corpus of the Lithuanian language in this paper. The corpus contains 4 speakers records of 731 distinct words. The duration of the records is near up to 1 hour. The corpus is built to cover all the most important features of common Lithuanian speech and is oriented towards different technologies. Corpus has it's own original sound search software *Tescal*.

The Speech Corpus was created before the researches at speech area for Lithuanian language were started. Creators of the corpus have decided, that corpus should be universal, medium size and detailed annotated. Such a corpus allows applications of different methods for Lithuanian language.

Structure of the corpus

Principles for choosing system of phonetic units

System of phonetic units was chosen as a base for corpus creation. The creation of the phonetic units system was based on primitive phonetic units and their main allophones. Phoneticians have determined such a units time ago, but they were not used for continuous speech analysis or speech synthesis. The principles for choosing phonetic units is described bellow. The *vowels* were discriminated from the *consonants*. No matter what approach it is (articulation, acoustical or functional), it's obvious that they belong to different sound classes.

Vowel based phonetic units

1. Vowels can be long or short. It is a very distinctive acoustical and functional feature. Short vowels: *a e i u o,* long: *a e i u o ė.*
2. All vowels can have accent or not. Long vowels can be accented in two types – with acute or circumflex. Short vowels without accentuation: *a e i u o;* accented short vowels: *a e i u o*. Long vowels without accentuation: *a e i u o ė;* long vowels with acute accent: *a e i u o ė;* long vowels with circumflex: *a e i u o ė.* So we have 28 vowels total.
3. Back vowels *o* and *u*, are used after soft consonants and became more similar to front vowels (differs their F2). The other distinction is between clear back vowels and little bit fronted back vowels. That forms 8 more units (considering accentuation).
4. It is better to treat *diphthongs* as individual phonetic units, than forming different vowels while joining them together. *Diphthongs* also can have acute accent (***a**i* ***a**u* ***i**e* ***e**u* ***e**i* ***u**i* ***u**o*)[1], circumflex (*a**i*** *a**u*** *i**e*** *e**u*** *e**i*** *u**i*** *u**o***) or be without accentuation (*ai au ie eu ei ui uo*). That brings additional 21 phonetic units.
5. *Vowel and consonant diphthongs* are treated as individual units too (e.g. *al am an ar el em en er ul um un ur il im in ir*). These *diphthongs* can be accentuated using the same manner as we had in *vowel-diphthongs*. Taking the accentuation into the account gives us 48 possible combinations. Then we should not forget that consonants may be hard or soft, so the number of previous combinations doubles. More over - *composite-diphthongs* having the consonant *n* used before consonants *k*, *g* change their place of articulation and due to this we add 24 units into our list of vowel-based phonetic units (aη aη’ **a**η **a**η’ aη aη’ eη eη’ **e**η **e**η’ eη eη’ iη iη’ **i**η **i**η’ iη iη’ uη uη’ **u**η **u**η’ uη uη’).

At this point we already have 175 vowel-based phonetic units (*composite-diphthongs* were included too as the majority of linguists believes those elements having the same functionality as the vowels - being the base for syllable). The words for the corpus were chosen so, that phonetic units (if it is possible) could be used at four positions: the beginning of the word, the middle of the word’s between voiceless consonants, the middle of the word’s between voiced consonants and the end of the word. Some of vowel-based units of course can not be used at all the positions we have mentioned. For example, diphthongs are rare at the end of the word.

If we would look at the phonetic features of the vowels even more precisely, it would be possible to combine 525 vowel-based elements.

Consonant based phonetic units

While choosing consonant based elements for initial analysis, primarily voiceless consonants were separated from voiced consonants. The boundaries between voiceless consonants and neighbouring vowels or voiced consonants can be set quite easy, so these units were taken without bigger part of vowel. The only feature that was taken into account was their softness: *c c’ ch ch’ f f’ k k’ p p’ s s’ sh sh’ t t’ č č’ x x’*. Seldom voiced consonants also were chosen without context of any vowels: *h h’ z z’ ž ž’ dz dz’ dž dž’*. So after that we have 28 units. While choosing words for corpus recordings, these elements were used at the beginning and the middle of a word. The words were chosen so, that hard consonants would be used before *a* or *u* and soft consonants before *e* or *i*. After counting down all possible positions we would have 80 elements.

[1] Bold means place of accentuation

Boundaries between neighbouring vowels and voiced consonants, liquids and nasals can be set as easy as it was with voiceless consonants. These elements were chosen as follows:

a) Consonants with big part of neighbouring vowel (*a e i u o ė*). These elements appear at the beginning, middle and the end of the example words.

b) Consonants without neighbouring vowels. Examples were chosen so that hard consonants would be used at the beginning of the word before long vowels *a*, *u*, *o*; middle of the word neighbouring other hard consonant. The words with soft consonants were chosen so that soft consonant would appear at the beginning of the word before long vowels e, ė, y, and middle of the word neighbouring another soft consonant.

At this point we have 72 consonants based phonetic units: *b b' ba be bi bo bu bė d d' da de di do du dė g g' ga ge gi go gu gė l l' la le li lo lu lė m m' ma me mi mo mu mė n n' na ne ni no nu nė r r' ra re ri ro ru rė v v' va ve vi vo vu vė.*

The consonant *j* should be mentioned separately. Elements *ja, je, ji, jo, ju, jė* were chosen for initial analysis. *J* consonant representing words were chosen according the same principle as voiced consonants: beginning, middle and end of word.

Altogether 106 consonant based phonetic units were added to system of our phonetic units. System of phonetic units is given as a list. List contains 275 lines, where each line represents each phonetic unit. Lines of the list are constructed as follows: transcribed phonetic unit; plain words representing particular phonetic unit; segmented and transcribed words. All phonetic units are represented by 731 word.

Table 1. One line from phonetic units list, given to phonetic unit „be“

be	begalvis, obelis, blogybe	be-g-al'-v-i-s, o-be-l'-i-s, b-l-o-g'-i-be

Recording the corpus

four speakers – 2 male and 2 female – were chosen to record phonetic units list. Records were made in a silent environment, but not in professional records studio; digital recording equipment and professional microphone were used for that matter.

The recording process was controlled by linguists. Special folder for each speaker's record was created and named using speaker's initials. Phonetic units tracks took 90 MB of memory, records were saved using PCM 44100 Hz 16 bit mono format.

File system of the corpus

The track was divided into 275 ranges according to each phonetic unit from the list. Each range contains all words matching one phonetic unit. Ranges are saved in separate "wav" files and named by the phonetic unit name. For example the phonetic unit "be" sound is saved as "fbe.wav"

An annotated corpus must contain not only phonetic unit records, but also phonetically transcribed words and information. The text were aligned to corresponding sounds, to their beginning and end. This information is saved to 275 annotated files - corresponding to each phonetic unit. Because the corpus was annotated using freeware software "Praat"

(<http://www.fon.hum.uva.nl/praat/>), is the reason for the annotation files being saved as the “Praat” file format “TextGrid”.

VMU universal isolated-words corpus contains more than 2000 such files.

Phonetic transcription of annotated text

While creating the VMU corpus it was very important to make the corpus intelligible for authors of other language corpuses. It is the reason why not only names of files, but also phonetic transcription of texts (annotation files) were encoded using the universal mode.

Most world languages have individual writing signs and phonetic transcription systems. For corpus creators and researchers it is important to have the possibility to compare phonetic systems of different languages and apply it to their personal research and software investigating other languages. There has been attempts towards creating a universal phonetic transcription system which would allow encoding language independent pronunciation features using ASCII codes. Unfortunately such universal system was not created and for the encoding of the VMU corpus a separate system was adapted and proposed. (see Table 2).

Table 2. Corpus adapted system: phonetic annotation using ASCII symbols

Phonetic units		Notation	Example
I.SOUNDS			
Vowels	Short vowels	Short vowels without accent marks *a, e, i, u* noted by common letters	a, e, i, u
		o without accent mark is noted special	o – oq
	Long vowels	Long vowels without accent marks *a, e, i, u* are noted by double letters	ą – aa, ę – ee, y, į – ii, ū,ų – uu
		ė without accent mark is noted special	ė – eh
Consonants	Consonants found in Latin alphabet	Noted by common letters	b, c, d, f, g, h, j, k, l, m, n, p, r, s, t, v, z
	2.2. Alveolar consonants, palatal *n* and *ch* 2.3. Soft consonants	Noted special Noted by „1“	č – ch, š – sh, ž – zh, ch – x, η - nq, k'– k1
II. PAUSES		Pauses between words, sentences, spontaneous breathes noted by w	
III. ACCENTS			
1. Grave		Noted by „4“ after main sound corresponding letter	à – a4

2. Circumflex	Noted by „3“ after main sound corresponding letter	ã – a3
3. Acute	Noted by „9“ after main sound corresponding letter	á – a9

How to annotate

Each sound recording is annotated with words, phonetic units and sound level. To do this word bands, phonetic unit bands and sound bands are created. Each band contains information about beginning and end times for a word, the phonetic unit boundaries of a word and phonemes boundaries. Experts can set segmentation boundaries for words, phonetic units or phonemes, see the sound spectrogram and listen to the recording. Such information is saved to the "TextGrid" files.

Validation and error documentation

While creating a corpus it's nearly impossible to avoid errors even less impossible to fix them so it's important to register these errors and save them in the documentation. Corpuses differ in character and errors differ accordingly. Several kinds of errors were found in validating the Lithuanian corpus. These were:
a) *Errors of the speaker*. The speaker goes over a standard text incorrectly. In these cases when words are missing or spoken words differ completely from the given text and don't represent the required phonetic units then the recordings were simply repeated. In other cases when the speaker changes the text in a minor way, new text is created which match the actual recording. Such kind of errors was scanned by the corpus authors and is called *inner validation*. After this validation stage individual texts for each speaker were created.

Pronunciation errors is the other kind of errors of the speaker, they appear when text is being read correctly, but the lengths of particular phonemes, accentuation is pronounced wrong or other phonetic errors are made. The process of finding such errors is called *outer validation*. Phonetic specialists find these kind of errors. This validation stage is crucial, but it fixes many pronunciation errors. Nearly 60 errors were fixed in the corpus we present.

b) *Encoding errors*. These sort of errors occur as simply mistakes, while wrong symbols are saved into annotation files. Such errors also can appear in the file names, but most of them are being found in the annotation texts. It is known that encoding mistakes appear in every fourth word. They might be found even after the correction.

c) *Segmentation errors* appear after the boundaries of elements are set at wrong place. This kind of errors are being fixed during both inner validation and outer validation, but they are hard to notice. Usually these errors are being found by researches. Such errors are documented for the resegmentation of the corpus.

Works based on the corpus

The initial experiments based on VMU Speech Corpus were started at 2001, when the primary version of the corpus were segmented. The purpose of the first researches done with the corpus was recognition of phonetic units using kNN, DTW algorithms. Later when the corpus was finally segmented, more researches were done. Widely applied technologies like MBROLA or HTK were practiced. We will overview the important researches from main areas of speech technologies that have been done with this corpus.

„MBROLA Software Suitability for Lithuanian Speech Synthesis“

The potential of MBROLA software in Lithuanian speech synthesis was investigated in this research. L.Žebrauskas presented the first examples of Lithuanian speech synthesis done with MBROLA. The MBROLA synthesis is based on the TD-PSOLA algorithm, which is designed for speech sound fusion and provides possibility to adapt length and pitch of the sounds to be synthesized. He has used the database of 1321 diphones obtained from the annotated VMU Speech Corpus. And found that the database was insufficient for good quality synthesis.

After this experiment the extension of the diphone database and formation of phoneme length and intonation models were included in the plans for the future work.

„The Lithuanian Speech Sounds Lengths Variation' Research“

According to I. Radziukynienė, it is essential to know the lengths of sounds and their variations with the rate of speech while developing Lithuanian speech synthesis system. She has used both – the linguistic rules found in various literature sources and means of statistical analysis of VMU Lithuanian speech corpus to create a model of Lithuanian speech sound lengths. The research has revealed that the lengths of sounds vary with the rate of speech changing from slow to medium, and from medium to fast, but the pause duration between words vary the most. Her initial hypothesis, stating, that the lengths of vowels vary and the lengths of consonants remain constant, when the rate of speech changes, was rejected.

Contextual Phonetic Units Intonation (Pitch) Models for Lithuanian Speech

S.Talandytė describes a specific feature of language component – phonetic units intonation. Intonation is known as curves variation of pitch. Phonetic units intonation is dissociated from influence of syllables, phrases and sentential intonation. Phonetic units intonation models in her research were created by approximating the variation of pitch by linear and square curves.

Investigation and Optimization of the Parameters for the Lithuanian HMM-based speech recognition system

G.Raškinis, D.Raškinienė present their ongoing work on the development of a Lithuanian HMM speech recognition system. Triphone single-Gaussian HMM MFCC-based speech recognition system was developed using HTK toolkit before this research. This system was designed to solve a speaker-independent ~750 distinct isolated-word recognition task. The authors solve the parameter optimization problem related to speech recognition system. Parameters related to feature extraction, such as upper frequency limit, number of mel frequency channels and MFCCs, cepstral mean normalization were investigated. 5 different parameters influencing HMM learning, such as: the sets of phonemes, 2 HMM prototypes, 2 different decision tree based triphone clustering scripts were investigated in their experiment as well. VMU isolated-word Speech Corpus was used in all their investigations. The results of parameter optimization reduced WER for their recognition system from 48-19% to 9-3%.

Acoustic Modeling of Transition Moments between Phonemes by Means of HMM

M.Štrimaitis presents Lithuanian speech recognition experiments, which are based on transition moments between phonemes models. Phoneme transition is defined as one phonemes end (from phonemes centre) and phonemes, which goes after it, beginning (to phonemes centre). Hidden Markov models were created for all phonemes transition moments. They were trained using Baum-Welsh algorithm. Viterbi algorithm was used for recognition using HTK tool. 4 speakers, 750 isolated words, 1 hour speech corpus was used for creating acoustic models. The author has investigated recognition accuracy relationship within several parameters: Mel Frequency Cepstral Coefficient number, frame length and upper frequency limit. He has obtained an average WER from 28 % to 36 %. The reason why the results were poorer than in speech recognition systems, according to M.Štrimaitis was – Hidden Markov Model methodology specialization for cases, when sign vectors were almost constants. He also notes that sign vectors changes in transition moments between phonemes were huge.

Software for Semi-automatic Detection of the Segmentation Points for Spoken Lithuanian

Construction of a set of candidate segmentation points is the first stage towards automatic annotation of speech corpora. The candidate set must satisfy the following requirements:

a) each expert-defined segmentation point is paired with one candidate in the set at least;

b) there are as few candidates as possible that are not paired with expert-defined segmentation points;

c) candidates and expert-defined segmentation points can be paired if their temporal distance is less than a given threshold.

J.Dailidaitė, Š.Talandytė present a system for constructing candidate sets satisfying the above listed requirements. They describe the structure of the system, the feature set used for speech analysis and algorithms for analyzing and evaluating temporal differences violating set requirements.

Second-level Features for Automatic Recognition of Segmentation Boundaries

D.Kuliešienė presents the system for automatic speech segmentation. The system is based on symbolic training methods and the segmentation boundaries are found using logical rules. These rules were gained from training the system to discriminate two classes of segmentation boundaries – artefacts and true segmentation boundaries. Each training example was represented by vector, it's components were values of contextual features (second-level features) for the candidate of segmentation boundary.

The system of the second-level features was constructed as a set of operators that describe relations between moments of time at training instances and parameter tracks (first-level features) of speech signal at given moment. The author describes operators and rules used in her research. According to her, together they constrain a grammar that is capable to generate language for second-level features. Such language was used to represent acoustical context of training examples.

The weaknesses of the corpus

The corpus we have presented is not suitable for continuous speech researches as only individual words are stored in it (for example, intonation of phrases or sentences can not be explored using it). The corpus is suitable for phonetic unit level researches, but not suitable for recognition at word level (for example, the corpus is too small to serve for building speech recognition system, because such systems usually make recognition at word level and their training requires many examples of the same word). So far there were no researches done for speaker recognition based on this corpus. This makes it hard to say if the corpus is suitable for recognition purpose.

Conclusions

- We think that the authors of the VMU isolated-word Speech Corpus have chosen a proper strategy - to build an universal annotated speech corpus. This medium size corpus was successfully applied for experiments with common technologies like MBROLA or HMM.
- The specifics of the pronunciation at beginnings and ends of the words influences individual words a lot. Broader researches require longer continuous speech fragments. The phonetic units of the corpus are presented at short phrases.
- Some limitations and lack of phonetic units were noticed during researches. For that the reason a new expanded version of corpus is going to be built.

References

[1] A.Raškinis, G.Raškinis, A.Kazlauskienė, VDU bendrinės lietuvių šnekos universalus, anotuotas garsynas (Universal annotated VMU Lithuanian speech corpus), Proceedings of the Conference "Informacinės Technologijos", KTU Kaunas, 2003, IX 28-34.

[2] L.Žebrauskas, A.Raškinis, M.Tamošiunaitė, "MBROLA" programines įrangos tinkamumo lietuvių šnekos sintezei tyrimas (Investigation of MBROLA Software Suitability for Lithuanian Speech Synthesis), Proceedings of the Conference "Informacinės Technologijos", KTU Kaunas, 2003, IX 25-27.

[3] I.Radžiukynienė, Lietuvių šnekos garsų ilgių kaitos, kintant šnekos tempui, tyrimas (The Lithuanian Speech Sounds Lengths Variation' Research, When the Rate of Speech Changes), Proceedings of the Conference "Informacinės Technologijos", KTU Kaunas, 2004, 214-216.

[4] A.Raškinis, S.Talandytė, Nuo konteksto nepriklausomų lietuvių šnekos garsų intonacijos (pagrindinio tono) modeliai ir jų tyrimas (Construction and Research of out of Context Phonetic Units Intonation (Pitch) Models for Lithuanian Speech), Proceedings of the Conference "Informacinės Technologijos", KTU Kaunas, 2004, 217-222.

[5] G.Raškinis, D.Raškinienė. Lietuvių šnekos atpažinimo sistemos, pagrįstos paslėptais Markovo modeliais, parametrų tyrimas ir optimizacija (Parameter Investigation and Optimization for the Lithuanian

HMM-based speech recognition system), Proceedings of the Conference “Informacinės Technologijos”, KTU Kaunas, 2003, IX 41-48.

[6] M.Štrimaitis, Perėjimo momentų tarp fonemų akustinis modeliavimas naudojant paslėptus Markovo Modelius (Acoustic Modelling of Transition Moments between Phonemes by Means of Hidden Markov Models), Proceedings of the Conference “Informacinės Technologijos”, KTU Kaunas, 2004, 240-244.

[7] A.Raškinis, J.Dailidaitė, Š.Talandytė, Bendrinės lietuvių šnekos garsynų segmentavimo taškų automatizuoto parinkimo ir tyrimo sistema (Computer System for Semi-automatic Detection and Investigation of Segmentation Points of Standard Spoken Lithuanian), Proceedings of the Conference “Informacinės Technologijos”, KTU Kaunas, 2004, 206-216.

[8] A.Raškinis, G.Raškinis, D.Kuliešienė, Antros eilės požymių sistema šnekos signalo segmentavimo taškų atpažinimui (System of second-level features for automatic recognition of segmentation boundaries), Proceedings of the Conference “Informacinės Technologijos”, KTU Kaunas, 2005, 294-298