

Slovenska baza izgovorjav z Lombardovim efektom - SiLSD

Damjan Vlaj, Aleksandra Zögling Markuš, Marko Kos, Zdravko Kačič

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko
Smetanova ulica 17, 2000 Maribor, Slovenija
{damjan.vlaj, sandra.zogling, marko.kos, kacic}@uni-mb.si

Povzetek

Članek predstavlja korake pri pridobivanju govornega materiala in postopke označevanja Slovenske baze izgovorjav z Lombardovim efektom (SiLSD – Slovenian Lombard Speech Database), katere snemanje se je začelo v letu 2008. SiLSD¹ je bila posneta na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Namen tega članka je opisati strojno platformo, ki je bila uporabljena za pridobivanje govornega materiala, opis poteka snemanja, predstavitev strukture baze izgovorjav in predstavitev orodja, ki je bilo uporabljeno za obdelavo baze izgovorjav. Baza izgovorjav zajema posnetke desetih slovenskih govorcev, od tega petih moških in petih žensk. Vsak govorec je posnel besedila osmih različnih korpusov, in to v dveh snemalnih sejah v razmiku vsaj enega tedna. Struktura korpusa je podobna strukturi, ki je bila uporabljena v bazi izgovorjav SpeechDat II. Posneto je bilo približno 30 minut govornega materiala na enega govorca in na eno snemalno sejo. Govorni material je bil ročno obdelan in označen z orodjem LombardSpeechLabel, razvitem na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Govorni material in pripadajoč označevalni material je shranjen na 10 DVD medijih (en govorec na enem DVD mediju).

Abstract

This paper presents the steps for acquisition of speech material and steps for annotation of Slovenian Lombard Speech Database (SiLSD), the recording of which started in the year 2008. The database was recorded at the Faculty of Electrical Engineering and Computer Science, University of Maribor. The goal of this paper is to describe the hardware platform used for the acquisition of speech material, description of recording scenarios, presentation of database structure and tools used for the annotation of SiLSD. The database consists of recordings of 10 Slovenian native speakers. Five males and five females were recorded. Each speaker pronounced a set of eight corpora in two recording sessions with at least one week pause between recordings. The structure of the corpus is similar to the SpeechDat II database. Approximately 30 minutes of speech material per speaker and per session was recorded. The manual annotation of speech material was performed with the LombardSpeechLabel tool developed at the Faculty of Electrical Engineering and Computer Science, University of Maribor. The speech and annotation material was saved on 10 DVDs (one speaker on one DVD).

1. Uvod

Namen članka je predstaviti strojno platformo, ki je bila uporabljena za pridobivanje govornega materiala, potek snemanja, strukturo baze izgovorjav in orodje, ki je bilo uporabljeno za obdelavo Slovenske baze izgovorjav z Lombardovim efektom (SiLSD – Slovenian Lombard Speech Database). Baza izgovorjav je bila posneta z namenom, da bi v posnetkih govora zajeli Lombardov efekt.

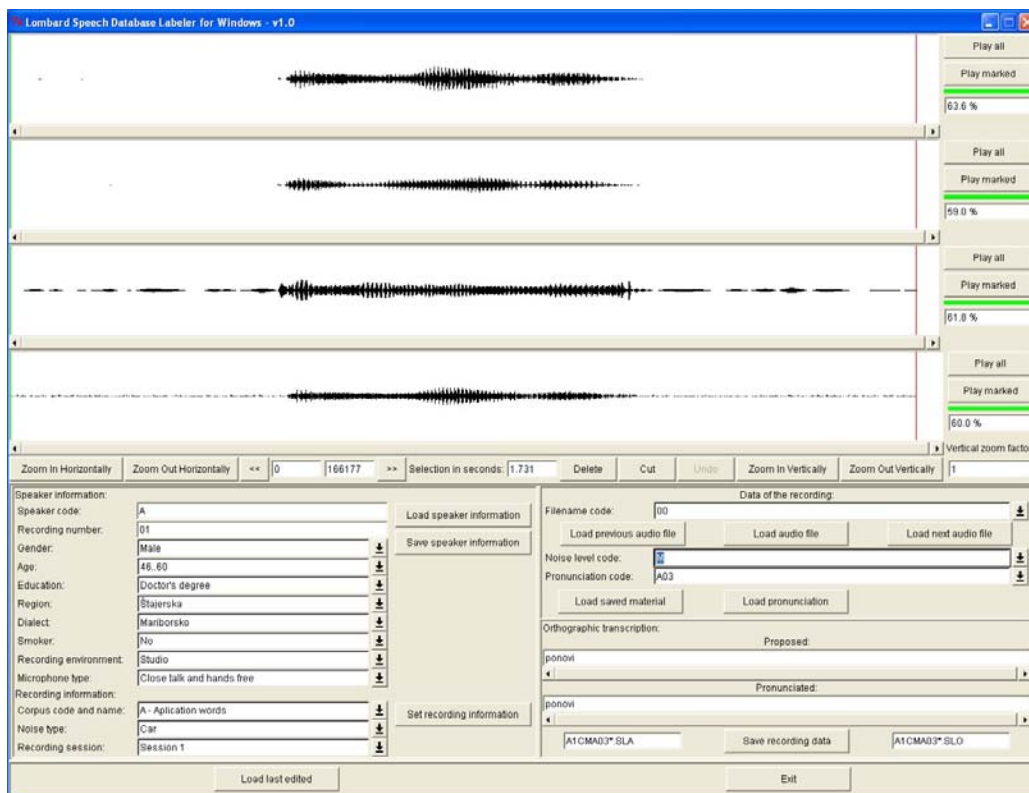
Lombardov efekt je bil prvič omenjen leta 1911, ko je Etienne Lombard (Lombard, 1911) odkril fiziološki efekt pri tvorjenju govora ob prisotnosti šuma. Lombardov efekt je pojav, pri katerem govorec poveča glasnost govora ob povečanju glasnosti šuma ozadja. Skozi zgodovino sta se pojavili dve interpretaciji Lombardovega efekta. Prva dokazuje, da je Lombardov efekt fiziološki avdio-fonetični refleks (Lombard, 1911). Druga razlaga pa temelji na domnevi, da so spremembe v govoru, ki jih povzročijo Lombardov efekt poledica slabše razumljivosti običajnega govora, ki ga govorec posluša v šumnem okolju (Lane in Tranel, 1971). Nekateri avtorji so tudi trdili, da lahko oba mehanizma prispevata k spremembam govornih karakteristik, kadar je govorec v šumnem okolju (Junqua, 1993) in tako povečata razumljivost govora. Ker povečanje glasnosti govora privede do sprememb karakteristik govornega signala, ima to pri avtomatskem razpoznavanju govora največkrat za posledico slabšo uspešnost avtomatskega razpoznavanja govora.

V procesu avtomatskega razpoznavanja govora ima postopek izločanja značilnik velik pomen in zelo vpliva na uspešnost sistemov avtomatskega razpoznavanja govora. Z uporabo standardnih postopkov izločanja značilnik, pri tem imamo v mislih uporabo mel-kepstalnih značilnik govornega signala in uporabo prikritih modelov Markova (Young in dr., 2000), se uspešnost avtomatskega razpoznavanja govora v študijskih razmerah približa 100 odstotkom (Hirsch in Pearce, 2000). Kakor hitro pa preidemo iz študijskega v naravno šumno okolje, uspešnost razpoznavanja upade glede na nivo razmerja med signalom in šumom (Hirsch in Pearce, 2000). Uspešnost avtomatskega razpoznavanja govora pa upade tudi zaradi sprememb karakteristik govornega signala, ki so se zgodile zaradi vpliva Lombardovega efekta na govorni signal. Vzrok za to so postopki izločanja značilnik in učenja prikritih modelov Markova, ki so izvedeni na bazah izgovorjav, ki niso posnete v okoljih, kjer je bil v govornem signalu prisoten Lombardov efekt.

Raziskave na področju Lombardovega efekta kažejo, da se Lombardov govor razlikuje od normalnega govora na več načinov. Glavne spremembe značilnosti Lombardovega govora so lahko vidne v zvišanju osnovne harmonske frekvence, v povečanju časovnega trajanja vokalov in v premiku formantov F1 in F2. Avtorja Hanley in Steer (Hanley in Steer, 1949) sta tudi ugotovila, da se hitrost govora zniža, ko je govor tvoren v šumnem okolju.

Baze izgovorjav posnete v realnem okolju zagotavljajo dragocen material za sisteme avtomatskega razpoznavanja govora, vendar je v primeru glasnejšega šumnega okolja

¹ Lastnik baze izgovorjav je podjetje SVOX, ki namerava ponuditi bazo izgovorjav preko organizacije ELDA/ELRA.



Slika 1: Programsko orodje LombardSpeechLabel za ročno obdelavo in označevanje govornega materiala.

(hrup v avtomobilu, govor v ozadju, ...) zaradi pomešanosti govora s šumom ozadja analiza prisotnosti Lombardovega efekta zelo otežena (Bořil in dr., 2006). Da bi bilo mogoče potrditi prisotnost Lombardovega efekta v govornem signalu, je potrebno analizirati govorni signal, ki vsebuje čim manj šuma v ozadju govora. Zato so Bořil in dr. posneli bazo izgovorjav, v kateri so želeli poudariti vpliv Lombardovega efekta v govornem signalu (Bořil in dr., 2006).

Z namenom nadaljevati raziskave na tem področju smo zasnovali snemalno okolje in izvedli snemanje baze izgovorjav SiLSD. Namen baze je omogočiti raziskave vpliva Lombardovega efekta v primeru različnih šumov okolja, različnih šumnih nivojev, ki jih v času snemanja sliši govorec in analizirati konsistentnost Lombardovega efekta glede na čas snemanja. Uporabili smo dva različna tipa šumov in izvedli snemanje pri dveh različnih nivojih šuma ozadja ter v dveh snemalnih sejah. S takšno bazo smo želeli omogočiti analize, ki bi potrdile, da povečanje nivoja šuma ozadja poveča vpliv Lombardovega efekta. Želeli smo tudi analizirati konsistentnost Lombardovega efekta znotraj dveh snemalnih sej, saj je med snemanjem posameznih sej s posameznim govorcem bilo vsaj en teden razmika.

V nadaljevanju tega članka, to je v drugem poglavju, bomo predstavili način pridobivanja govornega materiala, in opisali potek snemanja. V tretjem poglavju bomo predstavili način obdelave govornega materiala. V četrtem poglavju bomo predstavili strukturo baze izgovorjav SiLSD. V petem poglavju bomo podali zaključke.

2. Pridobivanje govornega materiala

Baza izgovorjav SiLSD je bila posneta v studijskem okolju. Vsak govorec je izgovoril nabor osmih korpusov

besedil v dveh snemalnih sejah z vsaj enim tednom razmika med snemalno sejo za posameznega govorca. Posneli smo približno 30 minut govornega materiala na enega govorca in na eno snemalno sejo.

Pri snemanju so bili hkrati uporabljeni prostoročni mikrofoni AKG C 3000 B, obustni mikrofoni Shure Beta 53 in dvokanalni laringograf EG2. Hkrati smo izvedli snemanje štirih kanalov:

- prostoročni mikrofoni,
- obustni mikrofoni,
- laringograf in
- snemanje šuma ozadja, pomešanega z govorom govorca, ki je bil predvajan na slušalke govorca med samim snemanjem.

Snemalna platforma je bila sestavljena iz zunanje zvočne kartice Audigy 4 PRO za štirikanalno zajemanje zvoka in mešalne mize Phonic MU244X. Zajemanje je bilo izvedeno pri frekvenci vzorčenja 96 kHz in 24-bitni linearni kvantizaciji.

Pri snemanju sta bila uporabljena dva tipa šumov in sicer hrup v avtomobilu in govor v ozadju. Šuma sta bila vzeta iz baze izgovorjav Aurora 2 (Hirsch in Pearce, 2000). Šuma sta bila normalizirana in predvajana na govorceve slušalke AKG K271.

Nivo predvajanega šuma je bil pred začetkom vsakega snemanja nastavljen na način, kot je bilo predlagano v (Bořil in dr., 2006). Zahtevan nivo šuma je bil nastavljen glede na učinkovito napetost zvočne kartice pri odprtih sponkah. Za doseg Lombardovega efekta smo izbrali nivo šuma 80 dB SPL² in 95 dB SPL pri navidezni razdalji med 1 in 3 metri.

² SPL je okrajšava za Sound Pressure Level (raven zvočnega tlaka).

Znotraj ene snemalne seje so bila izvedena tri snemanja:

- referenčno snemanje brez prisotnega šuma,
- snemanje pri 80 dB SPL in
- snemanje pri 95 dB SPL.

Med snemanjem posameznih besedilnih korpusov (besede, števila, povezane številke, stavki, ...) je bil narejen kratek premor, da se je lahko govorec prilagodil na okolje brez šuma. Daljši premor je bil narejen po snemanju vseh osmih besedilnih korpusov.

Med samim snemanjem je bilo vzpostavljeno sodelovanje med govorcem in snemalcem. Snemalec je slišal govor govorca z dodanim šumom. Pri tem je snemalec ocenil, ali je govor, ki ga sliši, razumljiv ali ne. Snemalec je preko LCD monitorja vzpodbujal govorca k bolj glasnemu govorjenju in ponovitvi izrečenega, če mu govor, ki ga je poslušal, ni bil dovolj razumljiv.

Baza izgovarjav zajema tako čiste posnetke brez dodanega šuma, kot tudi posnetke s šumom, ki so bili predvajani na slušalke govorca med snemanjem.

3. Obdelava govornega materiala

Govorni material je bil ročno pregledan, obdelan in označen s pomočjo programskega orodja Lombard-SpeechLabel (slika 1), ki je bilo razvito na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Programsko orodje je napisano v programskem jeziku Tcl/Tk/Tix, ki omogoča vizualno programiranje. Čeprav je bilo razvito na platformi operacijskega sistema Microsoft Windows, ga lahko brez večjih težav z manjšimi spremembami prenesemo na druge platforme operacijskih sistemov.

Okolje programskega orodja LombardSpeechLabel je razdeljeno v tri polja. Zgornje polje vsebuje prikaz štirih časovnih potekov signalov (prostoročni mikrofoni, obustni mikrofoni, laringograf in signal, predvajan na slušalke govorca), ki so bili zajeti med samim snemanjem baze izgovarjav. Vse posnetke je možno predvajati oz. poslušati s pritiskom na gumbe, ki se nahajajo na desni strani zgornjega polja. Spodnji del orodja je razdeljen na dva dela. Na levi strani se nahaja informacija o govorcju in snemanju. Na desni strani pa so podani dodatni podatki o snemanju in sama ortografska transkripcija posnetega signala.

Programsko orodje LombardSpeechLabel izvede tudi shrambo avdio materiala v končni format, tako da je signal vzorčen s frekvenco 96 kHz in kvantiziran s 16-bitno linearno kvantizacijo.

4. Struktura baze izgovarjav

Baza izgovarjav SiLSD vsebuje posnetke desetih slovenskih govorcev, od tega petih moških in petih žensk. Kot smo že omenili, je vsak govorec izgovoril nabor osmih korpusov besedil v dveh snemalnih sejah z vsaj enim tednom razmika med snemalno sejo za posameznega govorca. Struktura korpusa je podobna strukturi, ki je bila uporabljena v bazi izgovarjav SpeechDat II (Kaiser in Kačič, 1997). Več informacij o sami bazi izgovarjav bomo podali v naslednjih podpoglavjih.

4.1. Format avdio in opisne datoteke

Govorni signal je zapisan v avdio datotekah kot zaporedje otipkov kvantiziranih s 16-bitno linearno

A	Koda govorca (A-Z)
S	Koda seje (1-9) – uporabljeni samo 1 in 2
T	Koda tipa šuma: • R: brez šuma • C: hrup v avtomobilu • B: govor v ozadju
R	Koda snemanja: • N: snemanje referenčnega signala brez prisotnosti šuma • L: snemanje signala brez prisotnosti šuma • M: snemanje signala s prisotnostjo šuma pri nivoju šuma 80 dB SPL • H: snemanje signala s prisotnostjo šuma pri nivoju šuma 95 dB SPL
NNN	Koda korpusa besedil (A00 – Z99): A – aplikacijske besede B – povezane številke D – datumi I – izolirane številke N – naravna števila S – fonetično bogati stavki T – časi W – fonetično bogate besede
C	Koda snemalnega kanala: • 1: prostoročni mikrofoni • 2: obustni mikrofoni • 3: signal laringografa • 4: signal, ki je bil predvajan na slušalke govorca med snemanjem
LL	Dve črki jezikovne kode ISO 639
F	Koda tipa datoteke: O – datoteka z ortografsko označevalno vsebino A – datoteka z avdio vsebino

Tabela 1: Opis datotečne nomenklature.

kvantizacijo pri frekvenci vzorčenja 96 kHz. Otipki so shranjeni v surovem Intel formatu, brez kakršnekoli glave datoteke. Vsaka izgovarjava je shranjena v svoji avdio datoteki. Velikost datotek se razlikuje glede na besedilni korpus. Vsaka avdio datoteka ima pripadajočo opisno datoteko v opisnem formatu SAM s kodiranjem simbolov UTF-8.

4.2. Datotečna nomenklatura

Imena datotek sledijo datotečnemu zapisu ISO 9660 (8 + 3 znaki) glede na osnovni standard zgoščenk. Zaradi velike količine avdio materiala so bili vsi podatki shranjeni na DVD mediju.

Za datotečno nomenklaturu je bila uporabljena naslednja predloga:

A S T R N N N C . L L F

Datotečna nomenklatura je podrobno opisana v Tabeli 1.

4.3. Struktura map

Struktura map je sestavljena iz petih nivojev in je podana na naslednji način:

```

<database>
  <speaker>
    <session>
      <condition>
        <corpus>

```

<database>	Določena kot: <name><language code> i.e. LOMBSPSL kjer je: <name> LOMBSP in predstavlja Lombard Speech <LL> ISO 2-črki kode SL za Slovenian
<speaker>	Določen kot: SPK_<a> kjer <a> predstavlja naraščajočo črko abecede od A do Z. Ta črka je enaka prvi črki v imenu datoteke (glej poglavje 4.2).
<session>	Določena kot: SES_<s> kjer <s> predstavlja naraščajočo številko od 1 do 9. Ta številka je enaka drugi številki v imenu datoteke (glej poglavje 4.2).
<condition>	Določeni so trije tipi okoliščin: <ul style="list-style-type: none"> • REF: snemanje referenčnega signala brez prisotnosti šuma, • CAR: snemanje signala s prisotnostjo šuma hrupa v avtomobilu in • BABBLE: snemanje signala s prisotnostjo šuma govora v ozadju.
<corpus>	Določen kot: CORPUS_<c> kjer <c> predstavlja črko korpusa besedil: A – aplikacijske besede B – povezane številke D – datumi I – izolirane številke N – naravna števila S – fonetično bogati stavki T – časi W – fonetično bogate besede

Tabela 2: Struktura map za bazo izgovarjav SiLSD.

Struktura map je postavljena tako, da so posnetki vsakega govorca shranjeni na svojem DVD mediju. Vsak govorec je posnel svoj del baze izgovarjav v dveh sejah. V vsaki seji se nahajajo referenčni posnetki in posnetki, posneti pri dveh šumih okoljih, ki jih je slišal govorec. Vsako okolje vsebuje osem govornih korpusov.

Struktura map za bazo izgovarjav SiLSD je podrobneje podana v Tabeli 2.

4.4. Definicija kod korpusov besedil

V bazi izgovarjav smo definirali za vsak korpus besedil eno črko, ki določa korpus besedil, in dve številki, ki določata zaporedno številko posnetka v samem korpusu. Takšno označevanje je vneseno tudi v imena datotek, da bi uporabnik enostavno razbral iz imena datoteke, v kateri korpus besedil sodi njena vsebina. Definicija kod korpusov besedil je podana v Tabeli 3.

Predvsem pri izbiri fonetično bogatih besed in stavkov smo stremeli k čim bolj enakomerni pokritosti fonemov.

5. Zaključek

V članku smo opisali strojno platformo, ki je bila uporabljena za pridobivanje govornega materiala, podali smo potek snemanja, predstavili strukturo baze izgovarjav in orodje, ki smo ga uporabili za obdelavo baze izgovarjav SiLSD. Baza izgovarjav zajema posnetke desetih slovenskih govorcev, od tega petih moških in petih žensk.

Črka korpusa	Zaporedna številka	Vsebina korpusa besedil
A	00-29	aplikacijske besede (30 besed)
B	00-04	povezane številke (sekvenca 10 števk izgovorjena 5 krat)
D	00-04	datumi (5 datumov)
I	00-11	izolirane številke (12 števk)
N	00-04	naravna števila (5 števil)
S	00-29	fonetično bogati stavki (30 stavkov)
T	00-06	časi (7 časov)
W	00-49	fonetično bogate besede (50 besed)

Tabela 3: Definicija kod korpusov besedil.

Vsak govorec je posnel besedila osmih različnih korpusov in to v dveh snemalnih sejah v razmiku vsaj enega tedna med snemanji. Posneto je bilo približno 30 minut govornega materiala na enega govorca in na eno snemalno sejo.

Že med samim snemanjem so bile narejene nekatere analize govornega materiala, s katerimi smo potrdili prisotnost Lombardovega efekta v posnetem govornem signalu. Te analize so bile izvedene z opazovanjem sprememb osnovne harmonske frekvence, časovnega trajanja vokalov in s premiki formantov F1 in F2. V prihodnosti želimo izvesti poglobljeno analizo posnetega govornega materiala baze predvsem v smeri ugotavljanja vpliva različnih nivojev šuma ozadja na jakost Lombardovega efekta, odvisnost od vrste šuma ozadja in oceniti konsistentnost prisotnosti Lombardovega efekta v govornem signalu različnih govorcev.

6. Literatura

- Lombard, E. (1911). Le signe de l'elevation de la voix, *Annals maladiers oreille, Larynx, Nez, Pharynx*, 37: 101-119.
- Lane, H. in Tranel, B. (1971). The Lombard sign and the role of hearing in speech, *Journal of Speech and Hearing Research*, 14(4): 677-709.
- Junqua, J. C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers, *Journal of the Acoustical Society of America*, W(1): 510-524.
- Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V. in Woodland, P. (2000). *The HTK Book - Version 3.0*, Microsoft Corporation, ZDA.
- Hirsch, H. G. in Pearce, D. (2000). The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, *ISCA ITRW ASR'00 Proceedings*. Pariz, Francija.
- Hanley, T. in Steer, M. (1949). Effect of level of distracting noise upon speaking rate, duration and intensity, *Journal of Speech and Hearing Disorders*, 14(4): 363-368.
- Bořil, H., Bořil, T. in Pollák, P. (2006). Methodology of Lombard speech database acquisition: Experiences with CLSD, *Proceedings of the fifth Conference on Language Resources and Evaluation – LREC'06*, 1644-1647.
- Kaiser, J. in Kačič, Z. (1997). *SpeechDat Slovenian Database for the Fixed Telephone Network*, Univerza v Mariboru, Maribor, Slovenija.