Dragan Mićić i Mladen Bašić

Prepoznavanje govora pomoću žiroskopa

Ispitivane su mogućnosti prepoznavanja govora snimljenog žiroskopom, odnosno senzorom ugaone brzine. Osim same verifikacije ovakve metode prepoznavanja govora, ispitivano je kako položaj žiroskopa u odnosu na izvor zvuka i frekvencija odabiranja žiroskopa utiču na tačnost prepoznavanja ljudskog glasa na dobijenim snimcima. Najveća postignuta tačnost prepoznavanja iznosi 62% i dobijena je na bazi snimljenoj za muški glas pri korišćenju eksternog žiroskopa koji merenja vrši sa frekvencijom odabiranja 800Hz, dok najveća postignuta tačnost prepoznavanja u referentnom radu (Michalevsky et al*. 2014) iznosi 65%. Utvrđeno je da nije moguće registrovati reč ukoliko se žiroskop i izvor zvuka ne nalaze na zajedničkoj podlozi.*

Uvod

Osnovni cilj ovog rada jeste korišćenje senzora ugaone brzine odnosno žiroskopa za prepoznavanje govora snimljenog istim. Kako je zvuk mehanički talas, on stvara vibracije koje utiču na merenja ugaone brzine na žiroskopu. Zbog toga žiroskop indirektno meri zvučni talas koji do njega dolazi. Žiroskop u obliku elektronskog senzora postoji u svim modernim telefonima i koristi se za određivanje orijentacije telefona. Sve aplikacije imaju pristup žiroskopu, pa se otvara mogućnost njegovog korišćenja za snimanje i prepoznavanje govora korisnika bez njegovog znanja i dozvole.

U radu predstavljenom na konferenciji Black Hat 2014 (Michalevsky *et al*. 2014) audio snimci 10 ljudi koji su izgovorili 11 reči po 4 puta reprodukovani su na zvučnike. Ispred zvučnika se, na istom stolu, nalazio mobilni telefon koji je beležio očitavanja na žiroskopu frekvencijom odabiranja 200 Hz (eng. sampling rate). Analizom energije snimljenog signala, signal je segmentiran na reči i tišinu. Zatim je vršeno prepoznavanje izdvojenih reči zavisno i nezavisno od govornika. Prepoznavanje je vršeno pomoću klasifikatora maksimalne margine (eng. support vector machine), modela Gausovih smeša (eng. Gaussian mixture models) i dinamičkog vremenskog usklađivanja (eng. dynamic time warping). Najveća tačnost prepoznavanja dobijena je pri korišćenju metode dinamičkog vremenskog usklađivanja (u daljem tekstu DVU) zavisno od govornika i ona iznosi 65%. Naš rad za osnovni cilj ima reprodukciju i verifikaciju najuspešnijeg slučaja u referentnom radu, tj. postizanje tačnosti prepoznavanja od 65% na problemu prepoznavanja govora zavisno od govornika. Nakon toga, ispituju se okolnosti pod kojima žiroskop kao senzor zvuka ima najveću tačnost.

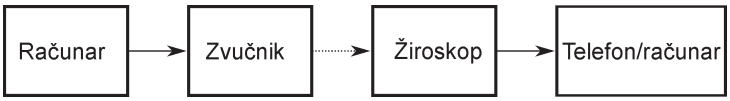
Po Nikvist-Šenonovoj teoremi odabiranja, za reprodukovanje signala maksimalne frekvencije f potrebna je frekvencija odabiranja $2f$. Shodno teoremi odabiranja, veća frekvencija odabiranja dovodi do veće tačnosti rekonstrukcije signala. Glavne karakteristike ljudskog glasa nalaze se u opsegu od 80 Hz do 1.1 kHz (Appelman 1967). Žiroskop telefona korišćenog u referentnom radu ima frekvenciju odabiranja 200 Hz, te po teoremi odabiranja iz snimaka napravljenih žiroskopom nije moguće tačno reprodukovati sve karakteristike ljudskog glasa.

Nazubljivanje (eng. aliasing) je efekat usled koga se odbirci odabirani frekvencijom f_s za signal frekvencije f registruju isto kao odbirci za signal frekvencije $|f - N f_s|$, gde je N prirodan broj.

Dragan Mićić (1998), Kruščica (Arilje), učenik 3. razreda Gimnazije „Sveti Sava" u Požegi

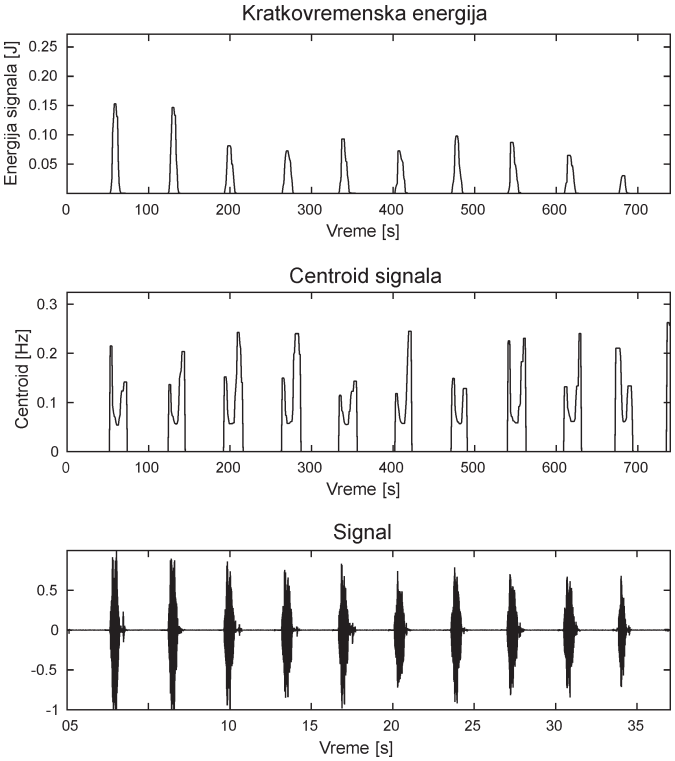
Mladen Bašić (1999), Vrnjci, Železnička 22, učenik 2. razreda Gimnazije Kraljevo

MENTOR: Marko Bežulj, Microsoft, ISP



Slika 1. Sistem za snimanje reči žiroskopom

Figure 1. Gyroscope recording mechanism



Slika 2. Algoritam za detekciju reči (Giannakopoulos 2014)

Figure 2. Word separation algorithm (Giannakopoulos 2014)

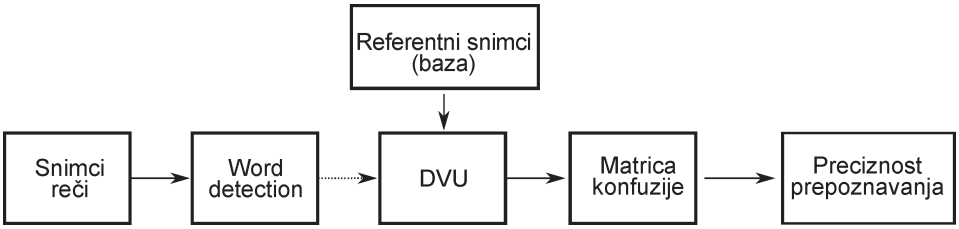
Ovi signali se nazivaju slikama signala frekvencije *f* i omogućavaju da postoje očitavanja signala čije su frekvencije veće od polovine frekvencije odabiranja. Prema tome pretpostavlja se da će tačnost prepoznavanja biti veća pri snimanju sa većom frekvencijom odabiranja na žiroskopu.

U eksperimentu opisanom u referentnom radu (Michalevsky i Boneh 2014) zvučnici i žiroskop se nalaze na istoj čvrstoj podlozi. Pretpostavka je da sa takvom postavkom, podloga na kojoj je žiroskop vibrira većim intenzitetom nego vazduh oko njega. Takođe, očekuje se da postoji preslušavanje (engl. crosstalk) između mikrofona i žiroskopa u samoj elektronici telefona, odnosno da će eksterni žiroskop imati manju tačnost prepoznavanja nego pri istoj frekvenciji odabiranja.

Metod

Po uzoru na referentni rad mikrofonom je snimljena baza od četvoro ljudi (dva muškaraca i dve žene) koji su svaku od 10 cifara izgovorili po 10 puta. Baza reči snimljenih mikrofonom, reprodukovana je na zvučnicima ispred kojih se nalazio žiroskop, čiji su tip, položaj i frekvencija odabiranja kasnije varirani (slika 1).

Iz dobijenih snimaka žiroskopa, pomoću algoritma za detekciju reči, isečeni su delovi u kojima se nalaze izgovorene reči. Korišćen je algoritam (Giannakopoulos 2014), koji poziciju reči u signalu detektuje merenjem vrednosti kratkovremenskih energija i centroida signala. Preklapanjem pikova funkcija energije i centroida, dobijeni su segmenti signala na kojima se nalaze izgovorene reči (slika 2). Nakon toga pojedi-

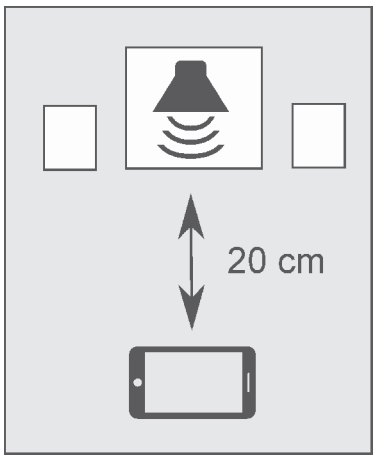


Slika 3. Sistem za detekciju reči

Figure 3. Word detection algorithm

načne reči se isecaju i čuvaju kao posebni signali. Tako isečene reči se prosleđuju u DVU algoritam (Muller 2007). Od 10 snimaka izgovora jedne cifre svake osobe, dva se koriste kao trening, a ostalih 8 kao test algoritma (slika 3). Svaku test cifru DVU algoritam upoređuje sa svim ciframa iz baze i određuje je kao onu sa kojom je imala najviše sličnosti. Rezultati prepoznavanja pojedinačnih cifara upisuju se u matricu konfuzije na osnovu koje se izračunava ukupna tačnost prepoznavanja.

Eksperiment 1. Kako bi se ispitao uticaj frekvencije odabiranja na tačnost prepoznavanja govora, u prvom eksperimentu snimci četvoro ljudi iz baze snimljene mikrofonom su ponovo uzorkovani na 100, 200, 800 i 8000 Hz i prosledili u sistem za prepoznavanje.

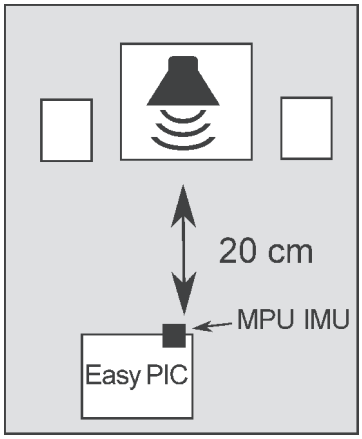


Slika 4. Postavaka eksperimenta 2

Figure 4. Experiment 2 setup

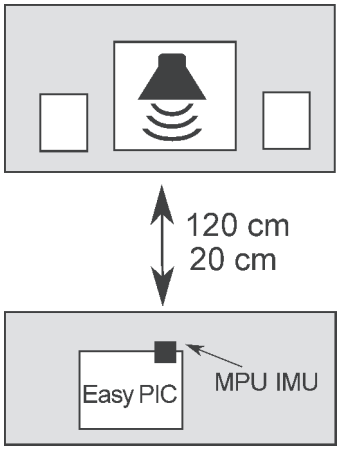
Eksperiment 2. Ovaj eksperiment predstavlja reprodukciju referentnog rada. Baza reči za jedan, muški glas, reprodukovana je na zvučnicima i snimana pomoću žiroskopa na mobilnom telefonu istom kao u referentnom radu (Samsung galaxy S3). Telefon se, kao i u referentnom radu, nalazio na istom stolu kao i zvučnici na razdaljini 20 cm (slika 4). Korišćena je Android aplikacija „Zmioskop" koja očitavanja sa žiroskopa zapisuje u fajl koji se, po prethodno opisanom sistemu (slika 3), prosleđuje algoritmu za detekciju reči, a potom u DVU algoritmu. Frekvencija odabiranja iznosila je 100 Hz, što je najveća frekvencija odabiranja koju je pomoću pomenute aplikacije moguće postići. U referentnom radu je, zahvaljujući novijoj verziji operativnog sistema telefona, postignuta frekvencija odabiranja 200 Hz.

Eksperiment 3. Baza reči reprodukovanih na zvučnicima snimana je eksternim žiroskopom. Kako bi merenja bila uporediva sa merenjima iz prethodnog eksperimenta, žiroskop je postavljen na istoj razdaljini (20 cm) od zvučnika koji se nalaze na istom stolu (slika 5). Korišćen je MPU IMU 6000 senzor koji je pomoću Easy PIC v7 razvojne ploče povezan na mikrokontroler (PIC 18f45k22). Mikrokontroler očitava merenja žiroskopa i USB UART konekcijom ih šalje na računar gde se čuvaju. Snimci su napravljeni za dva muška i dva ženska govornika frekvencijama odabiranja od 100, 200, 400 i 800 Hz – što je najveća frekvencija odabiranja koju smo sa eksternim žiroskopom uspeli da postignemo. Ovim eksperimentom želeli smo da ispitamo da li postoje preslušavanja (eng. crosstalk) između mikrofona i žiroskopa u elektronici mobilnog telefona. Ako postoje, prilikom korišćenja eksternog žiroskopa očekujemo lošije rezultate u



Slika 5. Postavka eksperimenta 3

Figure 5. Experiment 3 setup



Slika 6. Postavka eksperimenta 4

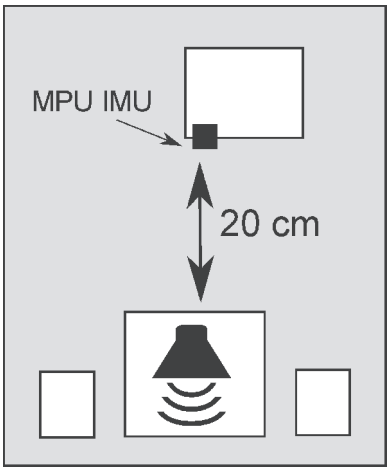
Figure 6. Experiment 4 setup

odnosu na one dobijene korišćenjem žiroskopa iz mobilog telefona.

Eksterni žiroskop može da beleži merenja frekvencijama odabiranja do 800 Hz. Zbog veće frekvencije odabiranja i jednostavnijeg postupka za ubiranje vrednosti merenja, u svim narednim merenjima korišćen je eksterni žiroskop

Eksperiment 4. Kako bismo ispitali da li se vibracije koje stvara zvučnik do žiroskopa prenose putem podloge ili vazduha, zvučinici i eksterni žiroskop su bili postavljeni na dva različita stola. Snimanje je ponovljeno za udaljenost između žiroskopa i zvučnika od 20 i 120 cm (slika 6). Korišćena je ista aparatura kao u eksperimentu 3, a snimci su napravljeni frekvencijom odabiranja 400 Hz samo za muški glas, čije je prepoznavaje u prethodnom eksperimentu bilo najtačnije.

Eksperiment 5. Kako bi se dodatno proverilo da li se vibracije koje stvara zvučnik do žiroskopa prenose vazduhom ili preko podloge, u ovom eksperimentu žiroskop je bio postavljen tako da se zvuk emituje u suprotnom pravcu od onog gde se žiroskop nalazi – iza zvučnika na razdaljini 20 cm (slika 7). Snimci su napravljeni istom aparaturom i pri istim parametrima kao u prethodna dva eksperimenta.

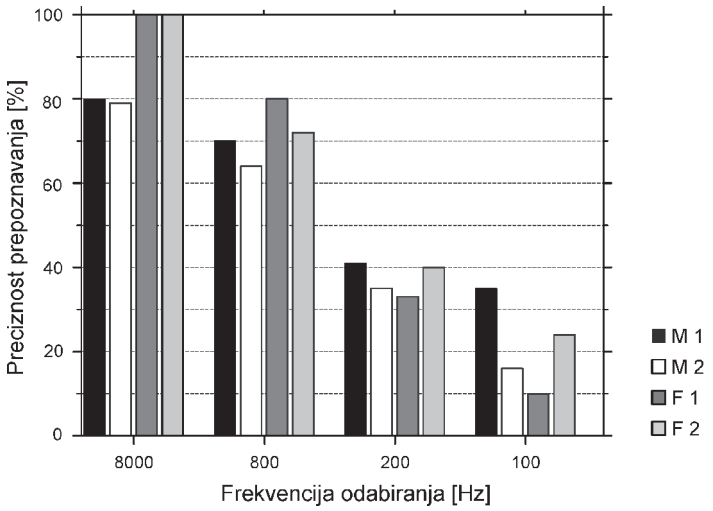


Slika 7. Postavka eksperimenta 5

Figure 7. Experiment 5 setup

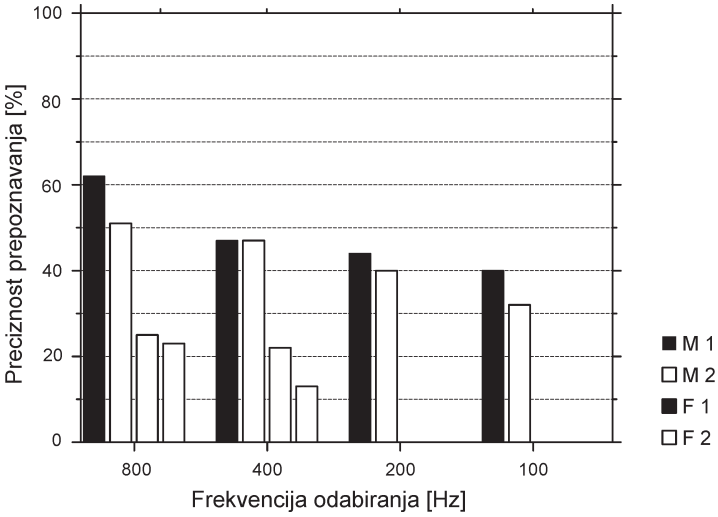
Rezultati i diskusija

Eksperiment 1. Tačnosti prepoznavanja govora za sistem sa jednim govornikom čiji je govor snimljen mikrofonom i ponovo odabiran na 100, 200, 800 i 8000 Hz je predstavljena na slici 8. Kao što je i očekivano, tačnost prepoznavanja opada pri manjim frekvencijama odabiranja. Takođe primećujemo da je pri nižim frekvencijama



Slika 8.
Rezultati eksperimenta 1:
M – muški glasovi
F – ženski glasovi

Figure 8.
Results of experiment 1:
M – male voices
F – female voices



Slika 9.
Rezultati eksperimenta 3:
M – muški glasovi
F – ženski glasovi

Figure 9.
Results of experiment 3:
M – male voices
F – female voices

odabiranja preciznost prepoznavanja veća za muške a pri višim za ženske glasove. Ove rezultate smatraćemo teorijskim maksimumom za algoritme koji koriste žiroskop za prepoznavanje govora.

Eksperiment 2. Na snimcima jednog govornika snimljenim žiroskopom iz mobilnog telefona pri frekvenciji odabiranja 100 Hz dobijena je tačnost prepoznavanja 36%. Za istog govornika teorijski maksimum tačnosti prepoznavanja pri frekvenciji odabiranja 100 Hz iznosi 35%, što je u redu veličine sa našim rezultatom.

Korišćenjem iste metode, ali pri frekvenciji odabiranja 200 Hz, u referentnom radu je postignuta tačnost prepoznavanja 65%. Zbog različite frekvencije odabiranja, rezultati našeg eksperimenta i rezultat referentnog rada nisu uporedivi. Međutim, tačnost prepoznavanja postignuta u referentnom radu je značajno veća od 41%, tj. od teorijskog maksimuma tačnosti prepoznavanja na snimcima odabiranim frekvencijom 200 Hz.

Eksperiment 3. Tačnost prepoznavanja zavisno od frekvencije odabiranja, na snimcima napravljenim eksternim žiroskopom za celu bazu reči, predstavljene su na slici 9. Može se primetiti da preciznost prepoznavana opada sa opadanjem frekvencije odabiranja. Vrednosti za tačnost prepoznavanja dobijena za muške glasove su

približne teorijskim maksimumima dobijenim u prvom eksperimentu. Na snimcima ženskih glasova na 100 i 200 Hz odnos signala i šuma manji je od 1, tako da nije bilo moguće detektovati sve reči, pa su ta merenja odbačena. Pretpostavljamo da je uzrok premalog odnosa signala i šuma to što ženski glas ima više elemenata visokih frekvencija te je, shodno teoremi odabiranja, za njihovo očitavanje potrebana viša frekvencija odabiranja.

Pri frekvenciji odabiranja 200Hz najveća tačnost prepoznavanja iznosi 40%. Pri istoj frekvenciji u referentnom radu dobijena je tačnost prepoznavanja od 65%. Za isti muški glas, sniman frekvencijom odabiranja 100 Hz, pomoću žiroskopa u mobilnom telefonu i eksternog žiroskopa dobijena je tačnost prepoznavanja 36% i 40% respektivno. Kako su ove dve vrednosti približne, može se zaključiti da ne postoje značajna preslušavanja u elektronici telefona.

Eksperiment 4. Na snimcima napravljenim žiroskopom koji se nalazi na drugom stolu na udaljenosti od 120 i 20 cm, odnos signal-šum je manji od 1. Algoritam za detekciju reči ne uspeva da detektuje, ili detektuje samo neke od reči. Ovaj rezultat potvrđuje hipotezu da se većina zvučnih talasa poteklih od zvučnika do žiroskopa prenosi preko podloge.

Ako bi ulogu zajedničke podloge mogla imati vilična kost, koja prenosi vibracije nastale u usnoj duplji do telefona prislonjenog na obraz korisnika, bilo bi moguće sa izvesnom tačnošću, prepoznati cifre koje korisnik izgovara prilikom razgovora mobilnim telefonom, što se navodi u referentnom radu (Michalevsky *et al*. 2014).

Eksperiment 5. Za snimke napravljene u eksperimentu 5 (kada je žiroskop postavljen iza zvučnika), tačnost prepoznavanja iznosi 38%. Rezultatom ovog eksperimenta dodatno je potvrđeno da se većina vibracija koje žiroskop očitava prenosi preko podloge.

Zaključak

Žiroskop mobilnog telefona i mikrofon imaju uporedivu tačnost prepoznavanja reči pri istoj frekvenciji odabiranja. Pri korišćenju žiroskopa na mobilnom telefonu nije postignuta ista frekvencija odabiranja kao u referentnom radu (Michalevsky *et al*. 2014), te on nije u potpunosti reprodukovan.

U eksperimentima 1 i 3 dobijena tačnost je značajno manja od vrednosti dobijenih u referentnom radu. Pretpostavljamo da je ova razlika u rezultatima nastala zbog značajno manjih baza reči u referentnom radu koje sadrže 5 puta manje snimaka po osobi od naših.

Kako eksterni žiroskop i žiroskop na telefonu imaju uporedive rezultate tačnosti prepoznavanja zaključeno je da nema značajnijeg preslušavanja u elektronici telefona.

Eksperimentima 4 i 5 je pokazano da žiroskop uspešno prepoznaje reprodukovane reči samo ako se izvor zvuka i žiroskop nalaze na istoj podlozi, u suprotnom žiroskop nije u mogućnosti da registruje reprodukovane reči. Ovaj rezultat potvrđuje hipotezu da se većina zvučnih talasa, poteklih od zvučnika, do žiroskopa prenosi preko podloge.

Kako je neophodno da izvor zvuka i žiroskop budu mehanički povezani i kako je opisani sistem namenjen za prepoznavaje reči zavisno od govornika, teško ga je praktično iskoristiti.

Zahvalnost. Zahvaljujemo se Lazaru Milenkoviću na kodiranju i ustupanju Android aplikacije „Zmioskop" korišćene za očitavanje merenja sa žiroskopa u mobilnom telefonu.

Literatura

Appelman D. 1967. *The science of vocal pedagogy*. Bloomington: Indiana University Press

Giannakopoulos T. 2014. A method for silence removal and segmentation of speech signals implemented in matlab. Dostupno na: http://cgi.di.uoa.gr/~tyiannak/Software.html

Michalevsky Y., Boneh D., Nakibly G. 2014. Gyrophone Recognizing Speech from Gyroscope Signals. U *Proceedings of the 23rd USENIX Security Symposium, August 20-22, 2014*. San Diego: USENIX Association, str. 1053-1067. Dostupno na: https://www.blackhat.com/docs/eu-14/materials/eu-14-Nakibly-Gyrophone-Eavesdropping-Using-A-Gyroscope-wp.pdf

Muller M. 2007. *Information retrieval for music and motion*. Heildelberg: Springer

Dragan Mićić and Mladen Bašić

Recognizing Speech with a Gyroscope

This paper presents approaches for recognizing a spoken text recorded by a gyroscope or by an angular rate sensor. Audio recordings of spoken numbers from 0 to 9 are reproduced from speakers while a gyroscope is positioned in front of the speakers. Sound is a mechanical wave and therefore it changes the angular velocity of the gyroscope used for measurement. The recordings are first separated by the number spoken. Recordings of individual numbers are then fed through the dynamic time warping algorithm. One part of the recordings is used for training and the other is used for testing the algorithm. In addition to the verification of the mentioned method for speech recognition, this paper deals with the examination of the parameters that affect this system, including the influence of the position of the gyroscope with respect to the sound source and the sample rate of the gyroscope. The system was tested using a smart phone integrated gyroscope, and also using an exterior electronic gyroscope. The highest accuracy of recognition was 62%, achieved with the male voice base while using an exterior gyroscope with a sampling frequency rate of 800 Hz.