МАТЕМАТИЧКА ГИМНАЗИЈА

МАТУРСКИ РАД

из предмета

Програмирање и програмски језици

на тему

Машинско препознавање емоција у говору

Ученик:
Анастасија Илић, IV_d

Ментор:
мс Милана Милошевић,
дипл. инж. ел. и рач.

Београд, мај 2018.

Садржај

1 Увод

Неке од најатрактивнијих области за истраживање у двадесет и првом веку јесу вештачка интелигенција и машинско учење. Идеја је научити машину да мисли. Међутим, људска бића су веома сложени организми и као такве, тешко их је моделовати. Проблеми моторике и говора су до неке мере савладани, међутим, као прави изазов намећу се емоције и емотивни говор.

Изражавање емоција преко говора је један од важнијих видова исказивања емоција. Познато је да је анализом говорног сигнала могуће утврдити више аспеката физичког и емоционалног стања као што су: године, пол, став, интелигенција и слично. Због тога, препознавање емоција у говору од посебног је значаја када је реч о комуникацији човек-машина као што су Интернет видео филмови, али и туторијали који се одржавају преко Интернета где одговор система зависи од уочене емоције код корисника. Терапеути могу анализом емоција у говору препознати депресију код човека која може на први поглед бити прикривена. Препознавање емоција у говору такође је коришћено приликом формирања апликације за центре корисничке подршке и мобилну комуникацију. Такође, једна од битних области где се примењује моделовање емоција у говору је синтеза говора. То је јако примењено у индустрији забаве - за анимиране филмове и ликове у игрицама. Синтеза емотивног говора неопходна је да би глас машине звучао природно и људски, а не равно.

Било да је у питању задатак препознавања говора, емоција у говору или говорника, креирање модела на основу снимљеног говора је од фундаменталног значаја. На основу доброг модела може се боље упознати теоретска природа сигнала, извора из ког сигнал потиче и слично. Сигнал говора је динамички сигнал и у зависности од задатка који је пред нас постављен потребно је да одаберемо адекватан модел којим ћемо овај сигнал описати. На пример, ако је у питању препознавање говорника, циљ нам је да модел говорника буде базиран на оним карактеристикама сигнала говора који не зависе ни од семантике, ни од расположења говорника, ни од околне буке или технологије. Коначно, најважнија примена модела за препознавање сигнала је то што нам помаже да схватимо и опишемо многе важне практичне системе: системе за предвиђања, системе за препознавање, идентификациони системе и сличне.

Моја мотивација за овај рад била је упознавање са машинским учењем кроз препознавање емоција у говору. Циљ овог рада био је упознавање са процесом обраде једног сигнала - од улазног звучног сигнала до класификатора сигнала по емоцијама. Централно место заузимају скривени Марковљеви модели. Осим моделовања емоција у говору, ови модели су се показали веома успешним и у другим областима као што је биоинформатика.

Овај рад састоји се из следећих целина: у другом поглављу објашњен је процес креирања базе емотивног говора. Након тога дат је преглед процеса издвајања карактеристике сигнала говора. У четвртом и петом поглављу теоретски су објашњени модели који се могу обучавати - у овом раду експеримент је вршен над скривеним Марковљевим моделима, па је у складу са тим тај алгоритам описан. У шестом и седмом поглављу објашњен је процес прављења система за препознавање емоција и тестирање једног таквог модела, као и резултати тог испитивања. На крају рада дат је закључак и преглед коришћене литературе.

2 База података говорних експресија, емоција и ставова - GEES

2.1 Технички подаци

Коришћена база емоционалног говора приликом тренирања модела је прва српска база емоционалног говора: база говорних експресија, емоција и ставова - GEES. То је уједно и једина званична база снимљеног говора на српском језику доступна за научно истраживање. Говорнике ове базе чине 3 мушкарца и 3 жене. Снимане емоције су: неутрална, бес, срећа, туга и страх. База се састоји из 32 изоловане речи, 30 кратких семантички неутралних реченица, 30 дугих семантички неутралних реченица и једног пасуса од 79 речи које изговара сваки говорник за сваку одређену емоцију. То укупно чини 2790 снимака емоционалног говора што је три сата говора.

2.2 Дизајнирање базе

Када се ради са базама емоционалног говора, битно је познавати процес постанка базе како би се омогућило што боље тренирање модела. Дизајнирање једне базе врши се у више етапа: одабир емоција, дизајн текста, снимање и процесирање, тестирање базе које укључује тест слушања, испитивање акустичних карактеристика и доношење закључака. Одабир емоција подразумева претходно познавање већ креираних база емоционалног говора како би се касније омогућило поређење резултата и потенцијално боље тренирање модела. Постоји више врста емоција, а одабране за снимање у српској бази података GEES биле су одглумљене, како би се испитивање одвијало у контролисаним условима. Снимање се одвијало на Факултету драмских уметности у Београду, у глувој соби. Сваки говорник сниман је одвојено како не би дошло до имитација емоција. Након снимања, извршен је тест слушања на основу кога је установљено колико човек добро препознаје емоције из ове базе. У табели 1 приказани су резултати теста слушања за GEES базу емотивног говора. Утврђена је тачност препознавања емоција од 95%.

Таблица 1: Резултати теста слушања за GEES базу емотивног говора

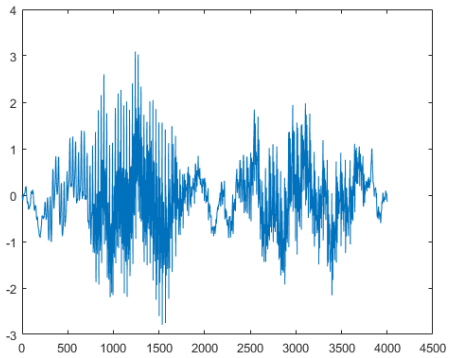
Говорник	Емоција		Одзив (препозната емоција у %)				
			Б	Н	Ср	Ст	Т
Сви	Тип емоције	Б	**96.06**	0.65	2.358	0.567	0.179
		Н	1.795	**94.67**	0.273	0.424	2.708
		Ср	2.302	0.89	**94.73**	1.606	0.312
		Ст	2.646	1.211	1.023	**93.33**	1.545
		Т	0.282	2.537	0.179	0.829	**94.04**

2.3 GEES база емоционалног говора

GEES база је доступна као део каталога Европске асоцијације за језичке ресурсе (ELRA) уз плаћање лиценце за истраживачке или комерцијалне сврхе. Ова база је једна од првих база одглумљеног емотивног говора. Са шест говорника, она је у рангу величине са већином осталих база одглумљеног емотивног говора. За сада је послужила за многа истраживања у области препознавања емоционалног говора, као и за израду овог рада.

3 Издвајање карактеристика говора

Како би се сигнал обрађивао, неопходно је пре свега из њега издвојити одређене карактеристике. У циљу тога потребно је познавати одређене теоријске карактеристике говорног сигнала, као и методе које се при издвајању тих карактеристика користе. На слици 1 приказан је сигнал представљен у временском домену.



Слика 1: Слика звучног сигнала - зависност амплитуде од времена

Описаћемо поступак обраде једног сигнала као и мотивацију за издвајање тачно одређених карактеристика.

3.1 Издвајање фрејмова из сигнала

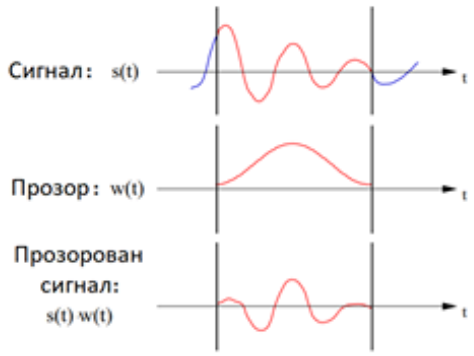
Сигнал се не обрађује у целости, већ се издвајају делови сигнала - фрејмови. Они су погоднији за обраду и даље операције са сигналом. Карактеристика једног сигнала је фреквенција одбирака F_s. То је учесталост са којом се издвајају дискретне вредности сигнала - тако се сигнал бележи у дигиталном облику. Фреквенција одбирака зависи од базе до базе. За GEES базу та фреквенција износи 44.1 kHz. Један фрејм се састоји од око 200 одбирака и дужине је 16 ms. Та дужина је уобичајена за намену препознавања емоција у говору.

3.2 Прозоровање

Приликом издвајања фрејмова из сигнала користи се метод прозоровања, конкретно, сигнал се множи са одређеном функцијом. Функција која се користи за прозоровање сигнала је Hamming window:

$$\omega(nT) = 0.54 - 0.46 \cos \frac{2\pi n}{N-1}$$

Прозоровање овог типа се користи како би се избегла спектрална дисторзија приликом фрејмовања сигнала. На слици 2 је шематски приказ прозоровања једног сигнала.



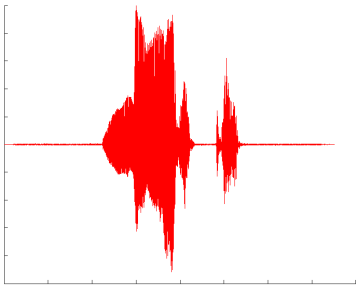
Слика 2: Прозоровање сигнала

3.3 Фуријеова трансформација

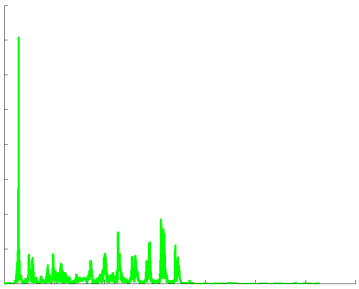
Фуријеова трансформација нам омогућава да израчунамо које фреквенције и на који начин се садрже у сигналу. Користимо је зато што је звучни сигнал лакше обрадити у спектралном домену него у временском. Сваки периодичан сигнал може се представити као збир синуса и косинуса.

$$g(t) = \sum_{m=0}^{\infty} a_m \cos \frac{2\pi m t}{T} + \sum_{n=1}^{\infty} b_n \sin \frac{2\pi n t}{T}$$

Када сигнал представимо као такав, амплитуде тих синуса и косинуса и фреквенције јединствено одређују сигнал. То су неке дискретне вредности које лако можемо да чувамо као тачну репрезентацију сигнала. Сигнал у временском домену је континуалан. Неопходне су нам дискретне вредности како бисмо га чували и касније представљали и због тога користимо Фуријеову трансформацију. У случају апериодичних сигнала сума прераста у интеграл. Овакве сигнале је и даље лакше описивати у фреквенцијском домену зато што већина реалних сигнала има ограничен фреквенцијски спектар. На сликама 3 и 4 приказани су, редом, сигнал пре Фуријеове трансформације - у временском домену, и након Фуријеове трансформације - у фреквентном домену.



Слика 3: Сигнал из GEES базе у временском домену



Слика 4: Сигнал из GEES базе у фреквентном домену

3.4 MFCC коефицијенти

Карактеристика сигнала коју користимо за препознавање емоција у говору јесу мел фреквенцијски кепстрални коефицијенти (MFCC). Они моделују начин на који људско уво чује. Такође, показали су се као супериорни у раличитим задацима обраде говора, па се радо користе у истраживањима. Поступак извлачења ових коефицијената из сигнала састоји се из следећих корака:

1. Фрејмован и прозорован сигнал се преводи на домен фреквенција по Фуријеовој трансформацији (FFT)

2. Спектар јачина се добија квадрирањем фреквенција које су претходно добијене Фуријеовом трансформацијом

3. Конструише се банка троугаоних филтара. Филтери су еквидистантни на мел-скали

$$H(k) = \begin{cases} 0 & k < f_{b_m} \\ \frac{k - f_{b_{m-1}}}{f_{b_m} - f_{b_m - 1}} & f_{b_{m-1}} \leq k \leq f_{b_m} \\ \frac{f_{b_{m+1}} - k}{f_{b_m+1} - f_{b_m}} & f_{b_m} \leq k \leq f_{b_{m+1}} \\ 1 & k > f_{b_{m+1}} \end{cases}$$

   где је m индекс филтра из банке троугаоних филтара, f_{b_m} су граничне тачке филтра, док је k k-та фреквенција од N у Фуријеовој трансформацији. Граничне тачке f_{b_m} су конструисане тако да деле мел скалу на m+1 подеок. Максимална фреквенција на мел скали одговара половини фреквенције одбирака, тј. $F_s/2$.

4. Филтри се потом трансформишу са мел скале на стандардну скалу по формули:

$$f_{mel} = 2595 \cdot \log_{10}(1 + \tfrac{f}{700})$$

   Уједно, ово је формула која представља везу између мел и линеарне скале.

5. Банка филтара се потом нормализује тако да је сума коефицијената за сваки филтар једнака 1. Овај корак даје коначан облик банци филтара.

6. Коришћењем филтара на спектру јачина добијамо мел спектар јачина.

7. На крају, MFCC коефицијенти се генеришу тако што се на логаритам мел спектра јачина примени дискретна косинусна трансформација.

За препознавање емоција у говору, у овом раду, коришћено је првих 13 коефицијената. Један вектор од 13 коефицијената се израчунава на основу једног фрејма. На тај начин, од једне реченице добијамо око 350 вектора од по 13 MFCC коефицијената. То израчунавање се врши над сваком реченицом, сваког говорника и за сваку емоцију. Уместо .wav фајлова након израчунавања MFCC коефицијената имамо .MFCC фајлове.

4 Модели за обучавање

У овом раду изучавана су два модела за обучавање за препознавање емоција у говору. Прво су изучаване Гаусове мешавине. Сложенији модел од овог јесу скривени Марковљеви модели. Гаусове мешавине представљају само једно стање у Марковљевим моделима и због тога их сматрамо поједностављеним моделом.

4.1 Гаусове мешавине

Гаусове мешавине (GMM) представљају параметарску пробабилистичку функцију густине вероватноће која се састоји из тежинске суме Гаусијана. Гаусове мешавине се најчешће користе као параметарски модел пробабилистичке расподеле континуалних величина које описују неки биометрички систем, као што су спектралне карактеристике сигнала.

Један Гаусијан представља функцију густине нормалне расподеле. Та функција густине за једнодимензиони вектор изгледа овако:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Међутим, чешће је случај да се посматра n-димензиони простор, па је зато један Гаусијан потребно проширити на n димензија, по формули

$$g(X|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{n/2} \mid \Sigma_i \mid^{1/2}} \exp\left\{-\frac{1}{2}(X-\mu_i)'\Sigma_i^{-1}(X-\mu_i)\right\}$$

где је X n-димензиони вектор који представља вектор неких карактеристика које моделујемо Гаусовим мешавинама, μ_i вектор средина и Σ_i дијагонална матрица коваријације.

Модел Гаусових мешавина је тежинска сума компонената Гаусијана дата формулом

$$p(X|\lambda) = \sum_{i=1}^{M} \alpha_i g(X|\mu_i, \Sigma_i) \qquad i = 1, \ldots, M,$$

где је X n-димензиони вектор, α_i су тежине Гаусијана, а $g(X|\mu_i, \Sigma_i)$ су управо Гаусијани као функције густине. За тежине важи $\sum_{i=1}^{M} \alpha_i = 1$.
Овај модел је параметарски, што значи да је циљ алгоритма кроз који он прође да се одреде ти параметри. У случају Гаусових мешавина и препознавања емоција у говору, тренинг алгоритам одређује параметре $\lambda = \{\mu_i, \Sigma_i, \alpha_i\}$. Почетни параметри се одређују на основу количине података над којом се модел обучава и на основу тога како се Гаусове мешавине користе у појединим биометричким системима.

Једна од предности коришћења Гаусових мешавина приликом моделовања јесте чињеница да се свака расподела произвољног облика може представити као сума појединачних отежињених Гаусових расподела. Гаусове мешавине представљају хибридни модел између класичне Гаусове расподеле и квантизиране расподеле (оне у којој су континуалне вредности представљене ограниченим сетом вредности).

4.2 Скривени Марковљеви модели

Скривени Марковљев модел (HMM) базира се на статистичким Марковљевим моделима у којима стања нису директно видљива.

Да би објаснили овај модел, најпре ћемо се упознати са појмом Марковљевих ланаца, а затим проширити теорију на скривене Марковљеве моделе.

4.2.1 Марковљеви ланци

Основа за рад са скривеним Марковљевим моделима јесу Марковљеви ланци.

Дефиниција 4.1. Нека је S пребројив скуп могућих стања $S = \{S_1, S_2, ..., S_n\}$ и нека је q_t стање у коме се систем налази у тренутку t. Овај низ стања q_i задовољава Марковљево својство ако важи

$$\mathbb{P}(q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, ...) = \mathbb{P}(q_t = S_j | q_{t-1} = S_i), \qquad 1 \leq i, j, k \leq n$$

и тада се назива Марковљев ланац.

Другим речима, предисторија не утиче на то у које ће стање систем прећи већ је од значаја искључиво његово тренутно стање.

Дефиниција 4.2. Марковљев ланац је временски хомоген уколико важи да је

$$\forall t, \mathbb{P}(q_t = S_j | q_{t-1} = S_i) = \mathbb{P}(q_2 = S_j | q_1 = S_i)$$

Коефицијенти транзиције су вероватноће преласка из једног у друго стање. Неопходне су за описивање Марковљевих ланаца и дефинишу се као:

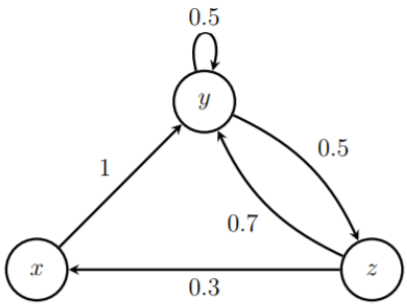
$$a_{i,j} = \mathbb{P}(q_t = S_j | q_{t-1} = S_i),$$

Коефицијенти транзиције задовољавају $\sum_{j=1}^{n} a_{i,j} = 1$
На основу коефицијената транзиције прави се стохастична или транзициона матрица. Димензије те матрице једнаке су кардиналности скупа дискретних вредности које случајне величине стања могу да узимају. Помоћу те матрице омогућује се лакше очитавање коефицијената транзиције.

Иницијална расподела је вероватноћа првог стања, пошто оно нема своје предстање. Дефинише се као

$$\pi_i(q_1) = \mathbb{P}(q_1 = i)$$

Дакле, како би се један Марковљев ланац у потпуности дефинисао неопходно је да он пре свега буде временски хомоген, а потом и да се одреде коефицијенти транзиције и иницијална расподела. На слици 5 приказан је пример Марковљевог ланца.



Слика 5: Пример Марковљевог ланца са коефицијентима транзиција и сетом стања $S = \{x, y, z\}$

4.2.2 Скривени Марковљев модел за обучавање

Код скривених Марковљевих модела имамо две врсте стохастичких процеса. Један процес омогућава транзицију кроз скуп стања, а други процес за задато стање генерише низ обсервација. Не можемо тачно утврдити у ком стању се систем налази, али на основу обсервација и статистичког модела можемо са одређеном вероватноћом претпоставити у ком стању се систем налази.

Оно што разликује скривене Марковљеве моделе од Марковљевих ланаца јесте појава обсервација уместо јасно дефинисаних стања и вероватноће појављивања појединих обсервација за свако стање понаособ.

Илустрација Скривених Марковљевих модела
Претпоставимо да је у некој соби невидљивој за посматрача дух и n урни. Посматрач не зна редослед тих урни. У свакој урни се налазе куглице различитих боја и нека постоји m различитих боја за куглице. Дух вади по једну куглицу из сваке урне и ставља је на покретну траку како би се она приказала посматрачу. Избор сваке урне из које вади куглице зависи искључиво од избора претходне урне. На основу приказаних куглица, посматрач може са одређеном вероватноћом да каже како су поређане урне. Тај закључак може извести управо преко вероватноћа налажења куглица одређене боје у свакој од урни.

4.2.3 Елементи скривених Марковљевих модела

Један скривени Марковљев модел дефинишу следећи параметри:

1. n је број стања у моделу. Скуп стања се обично означава са $S = \{S_1, S_2, ..., S_n\}$ За неке моделе постоји јасна представа шта та стања физички представљају. Међутим, када је препознавање емоција у говору у питању, та стања нам нису јасно позната, немају физичку интерпретацију.

2. m је број различитих обсервација која се могу реализовати из свих стања. Скуп обсервација се обично означава са $V = \{v_1, v_2, ..., v_m\}$. Код препознавања емоција у говору, за разлику од стања, обсервације имају своју физичку интерпретацију. Вектор са m MFCC коефицијената управо представља m обсервација.

3. Расподела вероватноћа транзиција $A = \{a_{i,j}\}$ где је

$$a_{i,j} = \mathbb{P}(q_t = S_j | q_{t-1} = S_i),$$

   Ове вероватноће се могу узети из било које расподеле, али је препоручљиво бирати оне које боље описују посматрани модел.

4. Вероватноће појављивања појединих симбола из појединих стања $B = \{b_j(k)\}$ где је

$$b_j(k) = \mathbb{P}(v_k, t | q_t = S_j)$$

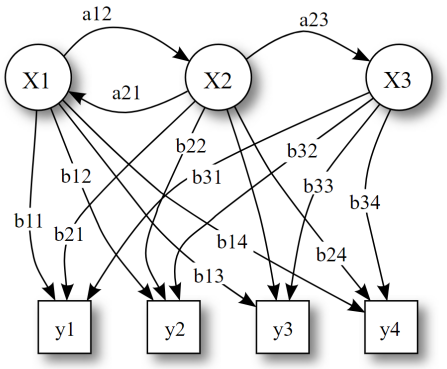
5. Иницијална расподела стања $p = \{\pi_i\}$ где је

$$\pi_i = \mathbb{P}(q_1 = S_i)$$

Дакле, параметри који описују HMM модел су n, m, A, B, p и на основу њега онда можемо формирати низ одговарајућих обсервација

$$O = O_1 O_2 ... O_T$$

где свако O_i означава једну од обсервација из скупа V. На слици 6 приказан је пример скривеног Марковљевог модела: $b_{i,j}$ су вероватноће појављивања одређених обсервација из одређених



Слика 6: Пример скривеног Марковљевог модела са коефицијентима транзиција и вероватноћом појављивања појединих симбола

стања, а $a_{i,j}$ су транзиционе вероватноће. Скуп обсервација је $V = \{y_1, y_2, y_3, y_4\}$, а скуп стања је $S = \{X_1, X_2, X_3\}$

Оно што је у овом раду било занимљиво јесте обрнути процес - од низа обсервација направити што бољи модел како би се та секвенца обсервација описала. Под прављењем модела подразумевамо налажење параметара

$$\lambda = (A, B, p)$$

који описују модел.

5 Три проблема скривених Марковљевих ланаца

Примена скривених Марковљевих ланаца је широка због природе реалних проблема. Приликом примене сусрећемо се са три проблема:

1. На основу датог низа обсервација израчунати вероватноћу да је модел $\mathbb{P}(O, \lambda)$ генерисао тај низ обсервација.
2. Ако нам је дат низ обсервација и модел израчунати секвенцу стања на основу ког је настала дата секвенца обсервација.
3. На основу низа обсервација одредити модел који је највероватније генерисао тај низ обсервација-максимизација $\mathbb{P}(O, \lambda)$

5.1 Први проблем

Нека је дата секвенца обсервација $O_1O_2...O_T$. Како би се израчунала вероватноћа појаве баш ове секвенце обсервација неопходно је пре свега одредити који редослед стања је могао да произведе ову секвенцу обсервација. Након тога је потребно одредити и са којом вероватноћом је свака секвенца стања генерисала дату секвенцу обсервација. Уколико фиксирамо секвенцу стања као $Q = q_1q_2...q_T$ можемо израчунати вероватноћу генерисања $O_1O_2...O_T$ као

$$\mathbb{P}(O|Q, \lambda) = \prod_{t=1}^{T} \mathbb{P}(O_t|q_t, \lambda)$$

Како смо већ увели вероватноће појављивања симбола, добијамо да је:

$$\mathbb{P}(O|Q, \lambda) = \prod_{t=1}^{T} b_{q_i}(O_i)$$

Такође, појава одређене секвенце стања такође има своју вероватноћу:

$$\mathbb{P}(Q|\lambda) = \pi_{q_1} \prod_{k=1}^{T-1} a_{q_k,q_{k+1}}$$

На основу формуле условне вероватноће добијамо да је:

$$\mathbb{P}(O, Q|\lambda) = \mathbb{P}(O|Q, \lambda)\mathbb{P}(Q, \lambda)$$

Комбинацијом свих горе наведених формула добијамо да је тражена вероватноћа:

$$P(O|\lambda) = \sum_{AllQ} \mathbb{P}(O|Q, \lambda)\mathbb{P}(Q, \lambda) = \sum_{q_1,q_2,...,q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1,q_2} b_{g_2}(O_2)...a_{q_{T-1},q_T} b_{q_T}(O_T)$$

Међутим, нумеричка сложеност оваквог израчунавања је јако велика. Због тога се уводе нови алгоритми: $Forward - Backward$ процедура. Она се заснива на динамичком програмирању. Уместо да рачунамо све одједном, у акумулатуру акумулирамо могућу вероватноћу до тренутка t. Forward променљива дефинише се на следећи начин:

$$\alpha_t(i) = \mathbb{P}(O_1O_2...O_t, q_t = S_i|\lambda)$$

и она представља вероватноћу дела обсервације до тренутка t, а у том тренутку стање S_i је активно. Индукцијом онда лако добијамо, са мањом нумеричком сложености, тражену вероватноћу. Аналогно можемо дефинисати и Backward променљиву:

$$\beta_t(i) = \mathbb{P}(O_{t+1}O_{t+2}...O_T, q_t = S_i|\lambda)$$

и она представља вероватноћу дела обсервације од тренутка t, а у том тренутку стање S_i је активно. Тражена вероватноћа такође се добија индукцијом, само што је сада у питању регресивна индукција. Од три проблема скривених Марковљевих модела, једино је овај екзактно решив; за све остале добијамо само оптималне вредности.

5.2 Други проблем

Нека је дата секвенца обсервација $O_1O_2...O_T$ и модел $\lambda = (A, B, p)$. Оно што се тражи јесте секвенца стања из којих је највероватније генерисан дати сет обсервација. Овај проблем је пробабилистички, другим речима, тачна стања не можемо добити. Проблем при решавању овог проблема је што не постоји универзалан критеријум оптималности. Због тога морамо да се одлучимо за одређени критеријум. Нека тај критеријум буде усвајање стања које је највероватније само по себи. За одређивање највероватнијег стања у неком тренутку t уводимо следећу функцију:

$$\gamma_t(i) = \mathbb{P}(q_t = S_i|O, \lambda)$$

Као што смо већ видели, динамички приступ решавању овог проблема је оптималнији. Функцију $\gamma_t(i)$ представљамо преко forward-backward функција:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\mathbb{P}(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{n} \alpha_t(i)\beta_t(i)}$$

Јасно је из формуле и смисла да важи:

$$\sum_{i=1}^{n} \gamma_t(i) = 1$$

На крају, за стање у тренутку t бирамо оно које је било највероватније на основу обсервације која се десила у истом том тренутку по формули

$$q_t = \arg\max_{1 \leq i \leq n}[\gamma_t(i)]$$

Постоје додатне оптимизације овог алгоритма, али о њима неће бити речи у овом раду.

5.3 Трећи проблем

Трећи проблем је најкомплекснији за решавање, али уједно и најбитнији зато што му је примена најшира. У реалности се обично сусрећемо са овим проблемом: имамо сет неких вредности и на основу њих правимо модел који их описује. То је суштина овог проблема. Имамо сет обсервација $O_1O_2...O_T$ и задатак је да направимо модел $\lambda = (A, B, p)$ за задату секвенцу. Као и у другом проблему, решење ће бити оно које је највероватније, али не и тачно решење. Идеја је да се изабере такав модел $\lambda = (A, B, p)$ који ће максимизовати $\mathbb{P}(O|\lambda)$. Процедура којом се одређује ова метода је итеративна и назива се $Baum - Welcher$-ова метода. Дефинишемо пре свега још једну функцију:

$$\xi_t(i, j) = \mathbb{P}(q_t = S_i, q_{t+1} = S_j|O, \lambda)$$

Користећи се свим функцијама које смо у претходна два проблема дефинисали добијамо:

$$\xi_t(i, j) = \frac{\alpha_t(i)a_{i,j}b_j(O_{t+1})\beta_{t+1}(j)}{P(O|\lambda)} = \frac{\alpha_t(i)a_{i,j}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{n}\sum_{i=1}^{n} \alpha_t(i)a_{i,j}b_j(O_{t+1})\beta_{t+1}(j)}$$

Параметар $\gamma_t(i)$ из претходног проблема је повезан са $\xi_t(i, j)$ по следећој формули:

$$\gamma_t(i) = \sum_{j=1}^{n} \xi_t(i, j)$$

Ако сумирамо по времену све $\gamma_t(i)$ и $\xi_t(i,j)$, добијамо очекивани број транзиција са i, односно са i на j, респективно.

$$\sum_{t=1}^{T-1} \gamma_t(i)$$

= очекивани број транзиција са i

$$\sum_{t=1}^{T-1} \xi_t(i,j)$$

=очекивани број транзиција са i на j
На основу свих наведених једначина, прави се апроксимација параметара који одређују модел:

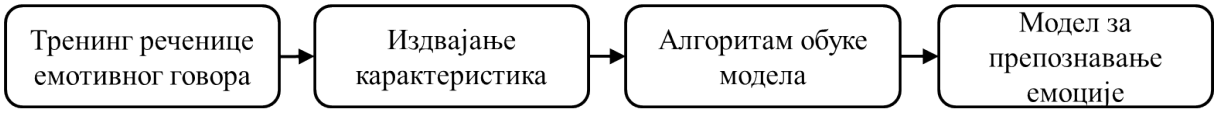
$$\overline{\pi_i} = \gamma_1(i)$$

$$\overline{a_{i,j}} = \frac{\sum\limits_{t=1}^{T-1} \xi_t(i,j)}{\sum\limits_{t=1}^{T-1} \gamma_t(i)}$$

$$\overline{b_j(k)} = \frac{\sum\limits_{t=1,O_t=v_k}^{T} \gamma_t(j)}{\sum\limits_{t=1}^{T} \gamma_t(j)}$$

6 Систем за препознавање емоција

До сада је имплементирано више софтвера за обраду говора. Од најстаријег PRAAT-а, до новијих, напреднијих софтвера као што су HTKtoolbox, OpenSmile, VoiceBox и слични. Како су модели за препознавање говора пробабилистички тј. није могуће одредити тачан модел, трага се за усавршавањем и надоградњом.
Елементи система за препознавање емоција су: база података, издвајање карактеристика, креирање модела емоција и тестирање. Шематски приказ система приказан је на слици 7.
База емотивног говора састоји се из .wav фајлова – то су управо звучни сигнали. У овом раду коришћена је GEES база коју смо описали у Глави 2.



Слика 7: Алгоритам прављења система за препознавање емоција

6.1 Издвајање карактеристика из реченица

Да би се вршила било каква тестирања или тренирања модела, неопходно је из звучних сигнала издвојити нумеричке карактеристике говора на основу којих обучавамо моделе и вршимо даљу анализу. Звучни сигнал је најпре подељен на фрејмове. У овом раду коришћени су фрејмови дужине од по 16 ms, са преклапањем од 9 ms. Што се тиче нумеричких карактеристика сигнала, у овом раду разматрани су и израчунавани MFCC коефицијенти. Процес поделе сигнала на фрејмове и израчунавање MFCC коефицијената описан је у Глави 3.

6.2 Избор тест и тренинг података

Зарад валидације система потребно је поделити податке на тренинг и тест реченице. Тренинг реченицама се обучавају модели, а тест реченице користе се у тестирању већ обученог модела. Постоје два начина поделе података: по говорницима и по проценту.

6.2.1 Издвајање реченица методом по говорнику

Приликом одвајања методом по говорнику, за тест реченице се издвајају све реченице једног говорника, док се на осталих 5 говорника тренира модел. Алгоритам тренирања модела се потом извршава 6 пута, са ротацијама говорника који се оставља за тестирање и као коначан резултат узима се средња вредност.

6.2.2 Издвајање реченица методом по проценту

Када је у питању издвајање реченица методом по проценту, за тест реченице узима се одређени број реченица за сваког говорника, док се на осталим реченицама врши тренирање модела. Метод паковања података за тренинг и тестирање у овом раду је био по проценту - од сваког говорника 70% реченица узимано је за тренинг, а 30% за тестирање модела.

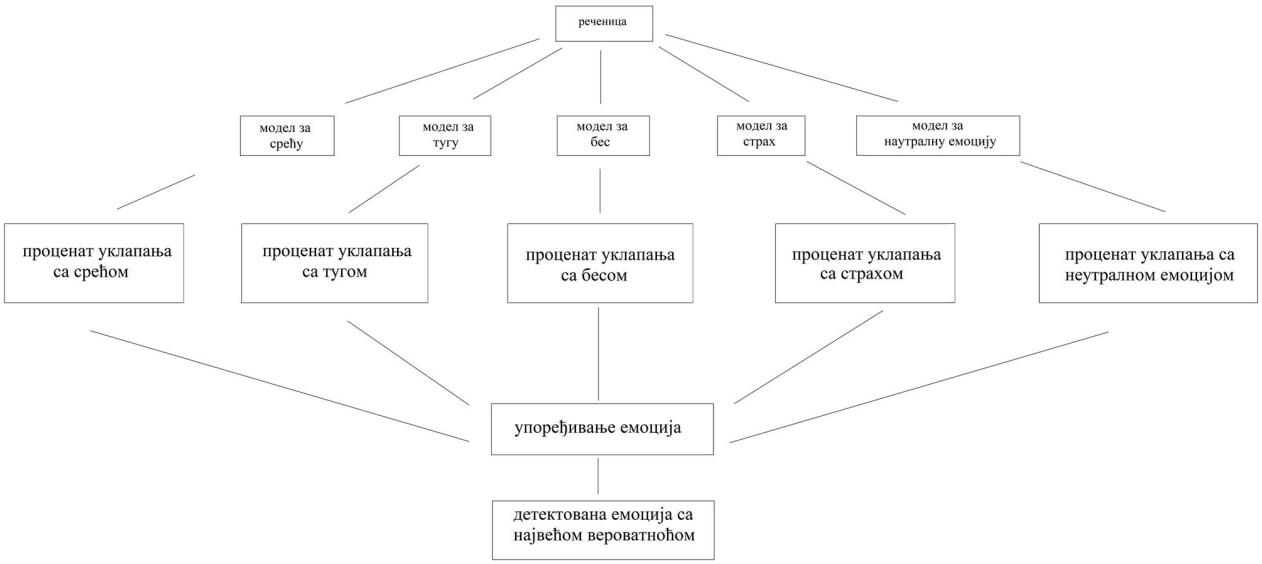
Одабир тренинг и тест реченица такође је јако битан део припреме података. Уколико приметимо да изражавање емоција једног говорника знатно одступа од осталих, много је боље употребити методу по проценту, а не остављати тог говорника за тестирање. Са друге стране, уколико нема значајних одступања, добро је тестирати и на потпуно новом сету података. Тиме постижемо боље испитивање модела, зато што ће сваки следећи пут модел добити потпуно нови сет података за који ће морати што боље да препозна емоцију.

6.3 Паковање података и припрема за тренинг алгоритам

Пошто постоје два начина поделе реченица, у складу са тим подаци се пакују у .mat фајлове. Овај корак је из практичних разлога убачен у софтвер. Циљ је да упакујемо све векторе MFCC коефицијената који су за једну одређену емоцију у један велики .mat фајл. У том фајлу су, дакле, све реченице предвиђене за тренинг од сваког говорника за једну емоцију. Овим смо оптимизовали учитавање података у тренинг алгоритам.

6.4 Тест алгоритам

Циљ тест алгоритма је препознати која емоција се највероватније крије у реченици коју тестирамо. Уколико се један вектор састоји из 13 MFCC коефицијената, говоримо о 13-димензионом простору, где су MFCC коефицијенти кординате једне тачке. За једну реченицу издвојили смо око 350 тачака. Резултат који даје модел на основу 350 тачака које припадају једној реченици јесте вероватноћа да реченица коју смо убацили у модел за, рецимо срећу, окарактерисана срећом. Поступак поновимо за свих пет модела емоција и на крају поредимо проценте које смо добили. Уколико је проценат највећи код модела који је истрениран за тугу, претпоставља се да дата реченица испољава тугу као емоцију. На слици 8 шематски је приказан алгоритам тестирања модела.



Слика 8: Тест алгоритам

6.5 Тренинг алгоритам

Циљ тренинг алгоритма је направити што бољи модел тако што се одреде коефицијенти A, B, p. Модел који ми користимо је скривени Марковљев модел. Креирамо по један модел за сваку емоцију.
Тренинг алгоритам се одвија у више етапа: иницијализација, тренирање модела, обучен модел.

6.5.1 Иницијализација

Параметри модела се најпре морају поставити на неке иницијалне вредности како би тренинг започео на неком моделу. У овом раду почетни параметри су изабрани случајно и потом су нормализовани. Ово је могуће зато што би алгоритам теоретски требало да конвергира ка неким вредностима, па иницијалне вредности нису битне.

6.5.2 Тренирање модела

Велика матрица свих MFCC коефицијената које смо претходно направили као један .mat фајл се пропушта кроз алгоритам. Почетне вредности модела који улази у алгоритам су коефицијенти A_{old}, B_{old}, p_{old}. Након проласка кроз једну итерацију тог алгоритма добијамо нове, поправљене коефицијенте A, B, p. Следећи корак је провера колико вероватноће које се рачунају као и у тест алгоритму на основу коефцијената одступају једна од друге. Уколико је то одступање мање од неког унапред одређеног прага, тренинг алгоритам се завршава и ми имамо истренирани модел за неку емоцију. Међутим, ако то није случај, нове вредности коефицијената сада постају старе и извршава се још једна итерација алгоритма и цео поступак се понавља. Могуће је да у једном тренутку не дође до конвергенције одступања, већ одступање почне да флуктуира између две вредности. Како би се избегло упадање алгоритма у бесконачну петљу, потребно је и ограничити број итерација након ког сматрамо да је модел истренирaн.

6.6 Обучени модел

Након извршавања тренинг алгоритма имамо модел за сваку од емоција - сваки модел једнозначно одређују добијени параметри A, B, p. Дакле, излаз из симулације се састоји из:

1) обрађених података (сигнали у облику .MFCC)

2) модела (пет модела за пет емоција који су резултат тренинг алгоритма)

3) резултата

Резултати се добијају када моделе које смо креирали тестирамо коришћењем тест реченица. Оно што добијамо као резултате јесте колико се која реченица уклапа у сваки од модела и на основу тога можемо да утврдимо која емоција је у питању. Овакав модел је спреман за препознавање емоција у новим реченицама и даља истраживања у области препознавања емоција у говору.

7 Резултати експеримента

Експерименти су извршени у програмском окружењу SEBAS који је развијан у оквиру докторских студија на Електротехничком факултету у Београду. Аутори окружења су Милана Милошевић и Жељко Недељковић.
Одрађена је обрада сигнала од самог почетка, преко паковања и тренирања модела до коначног тестирања.
За тренинг модел користили су се скривени Марковљеви модели.
Тренинг алгоритам је пуштен на српској бази емотивног говора GEES која је коришћена као основна база.
Резултати су приказани у табели кофнузије на слици 9. У таблици конфузије (слика 9) видимо

Време симулације: 26-Мај-2018 11:56:52

Параметри:

Класификатор	Скривени Марковљев модел
Карактеристике	MFCC коефицијенти
База	GEES
Обука	поценат = тест
Тест	Проценат
Укупна успешност	68.40 %

Матрица конфузије:

	неутрално	бес	срећа	туга	страх
неутрално	87.76	16.33	14.29	6.00	3.77
бес	4.08	57.14	59.18	0.00	13.21
срећа	2.04	20.41	22.45	0.00	0.00
туга	4.08	0.00	0.00	90.00	0.00
страх	2.04	6.12	4.08	4.00	83.02

Слика 9: Резлтати тренинга и тестирања на GEES бази методом по проценту

да највећи степен препознавања емоције имамо код туге, неутралне емоције и страха. Код среће и беса видимо да је резултат често био њихово међусобно мешање. То објашњавамо природом људских емоција. Наиме, и срећа и бес су екстремне емоције: повишене фреквенције, гласније, јаче емоције. Можемо закључити да MFCC коефицијенти нису довољно добри да разликују ове две емоције.

Ако изузмемо мешање беса и среће, овај систем за препознавање емоција показује висок степен препознавања. За остале три емоције, неутралну, тугу и страх он износи редом 87.76%, 90.0% и 83.02%. У GEES бази конкретно над овим подацима степен људског препознавања емоција био је око 95%. У поређењу са тим укупан резултат од 68.40% доста одступа од очекиваног. Међутим, генерално је степен људског препознавања емоција око 70%, па овај модел са MFCC коефицијентима можемо сматрати релативно успешним.

8 Закључак

Главни циљ овог матурског рада је био да представи теоријски и практично процес обраде сигнала до тренирања и тестирања модела за препознавање емоција у говору. Одабрано је препознавање емоција у говору зато што је то један од најбитнијих сегмената вештачке интелигенције и као млада област пружа доста могућности за напредак.

Приликом израде самог рада коришћена је сва доступна литература. Такође, за тренинг и тестирање модела коришћена је српска база емотивног говора GEES. Простор постоји за истраживање и тестирање на другим базама језика, као и на базама спонтаног говора уместо одглумљеног.

У поређењу са другим структурама моделима, као што је модел Гаусових мешавина, скривени Марковљеви модели односе апсолутну победу. Већина појава која нас окружује је пробабилистичка. Обично не можемо одредити тачну секвенцу стања кроз који неки систем пролази. Због тога модел као што је скривени Марковљев модел има предност у односу на остале детерминистичке моделе.

Као једна динамичка структура, скривени Марковљев модел показао се као модел који доста прецизно моделује људско препознавање емоција, са успешношћу од $68,40\%$. У области вештачке интелигенције, пожељно је превазићи и тај проценат и на томе се ради. До тада, овај модел може да буде од користи јер је на нивоу препознавања људских емоција.
Такође, један од битних параметара у експерименту били су MFCC коефицијенти. Како мешање само две емоције не може бити грешка модела, закључили смо да ови коефицијенти са мање успешности разликују бес и срећу.

Закључак рада је да је задатак препознавања емоција комплексан и зависи од више елемената. Ти елементи су: начин на који дефинишемо емоцију, говорници на основу чијег гласа се креира база података, избор реченица за тренирање и тест, избор карактеристика сигнала говора на основу којих се врши моделирање и њихових параметара и на крају избор модела који ће бити употребљен и његових параметара.

8.1 Стечено знање

Приликом израде матурског рада сам се детаљније упознала са концептом машинског учења. Научила сам више о итеративним алгоритмима који се користе у машинском учењу. Упознала сам се са програмским окружењем Matlab и предностима које оно има при обради сигнала. Што је најбитније, остварила сам циљ овог матурског рада, а то је да се у потпуности упознам са принципима и методама обраде сигнала и њихове класификације на основу емоције исказане тим говорним сигналом.

8.2 Даљи рад

Многи модели нису обрађени и многи коефицијенти нису разматрани у овом раду. Степен препознавања емоција преко машине је на завидном нивоу, али још увек није достигао људски ниво препознавања емоција. У даљем раду бих се посветила проналаску што бољих модела и коефицијената у циљу повећавања степена препознавања емоција у говору.

8.3 Захвалност

На крају бих желела бих да изразим захвалност

- **Милани Милошевић** – мојој менторки, докторандкињи Електротехничког факултета у Београду, која ми је значајно помогла при изради матурског рада, пре свега несебичним издвајањем времена и свеопштом присутности у овом раду. Дала ми је низ сугестија, критика и савета и увек је била максимално на располагању за све што ми је требало и са великим стрпљењем објашњавала све што ми није било јасно и подстицала ме да радим вредно и предано.

- **Филипу Марићу** – професору који је индиректно утицао на настанак овог рада, захваљујући коме сам значајно напредовала на пољу програмирања и разумевања алгоритама

- Свим професорима – који су кроз моје школовање имали утицај на развијање мог информатичког, али и свеукупног знања

- Свима осталима – који нису поменути овде, али су имали значајан утицај на развој мог рада

9 Литература

[1] BAUM, L. E., PETRIE, T., SOULES, G., & WEISS, N. (1970) *A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The annals of mathematical statistics, 164-171.*

[2] LAWRENCE R. RABINER, fellow, IEEE (1989) *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*

[3] LAWRENCE R. RABINER, BIING HWANG JUANG (1993) *Fundamentals of Speech Recognition*

[4] Slobodan T. Jovičić, Zorka Kašić, Miodrag Đorđević, Mirjana Rajković (2004) *Serbian emotional speech database: design, processing and evaluation*

[5] Željko Nedeljković, Milana Milošević, Željko Đurović *Baseline system for speaker recognition - parametar analysis*

[6] Douglas Reynolds, MIT Lincoln Laboratory *Gaussian Mixture Models*

[7] CHRISTOPHER J.C. BURGES *A Tutorial on Support Vector Machines for Pattern Recognition*

[8] Milana Milošević, Željko Nedeljković, Željko Đurović *SVM Classifier for Emotional Speech Recognition in Software Environment SEBAS*

[9] Milana Milošević, Željko Đurović (2015) *Challenges in Emotion Speech Recognition*

[10] Milana Milošević, Željko Đurović (2015) *Emotion Feature Extraction for Emotion Classification from Speech Signal*

[11] Machine learning, `https://en.wikipedia.org/wiki/Machine_learning`
Последњи приступ сајту: 26.05.2018.

[12] Hidden Markov Model, `https://en.wikipedia.org/wiki/Hidden_Markov_model`
Последњи приступ сајту: 26.05.2018.