Универзитет у Београду – Електротехнички факултет

МАСТЕР РАД

на тему

Синтеза видео записа на основу говорног сигнала употребом рекурентних неуралних мрежа

кандидат:
Вељко Шешељ, 3253/2018

ментор:
доц. др Предраг Тадић

август, 2020.

Захвалница

Садржај

Списак слика

Увод

1 | Обележја говора

Први корак у синтези лажног видеа из говора је одређивање карактеристичних обележја говорног сигнала. Компонентне аудио сигнала треба да су довољно добре да носе лингвистичку компоненту, али и да су отпорне на позадински шум и на остале сметње. Једна од основних претпоставки везаних за обраду говорног сигнала јесте да се говор може приказати као излаз линеарног, временски променљивог система, чија се својства споро мењају са временом. То води ка основном принципу анализе говора који каже да ако се посматрају довољно кратки сегменти говорног сигнала, да се тада сваки сегмент може моделирати као излаз линеарног, временски инваријантног система [1]. Стога се кратки сегменти говора могу описати конволуционом једначином

$$s(t) = e(t) * \theta(t) \tag{1.1}$$

при чему $s(t)$ представља резултатни говорни сигнал, $e(t)$ представља побудну ваздушну струју (екситацију) и $\Theta(t)$ импулсни одзив органа говорног тракта. У рачунарској обради говора, сигнале је згодније посматрати у дискретном домену

$$s[n] = e[n] * \theta[n] \tag{1.2}$$

Зависно од типа екситације, говорни гласови (фонеми) се могу поделити у три дистинктне категорије:

- звучни гласови
- фрикативи (безвучни гласови)
- плозиви

Код звучних гласова, ваздух прелази преко затегнутих гласних жица које почињу да вибрирају релаксираним осцилацијама, производећи квази-периодичне четвртке које ће побудити вокални тракт. Типични представници звучних гласова су самогласници. Побуда која производи фрикативе је окарактерисана широким спектралним садржајем као што је случајни шум (глас „ш”). Плозиви настају побудом која је високог интезитета и кратког трајања (попут Дираковог импулса) (гласови „б”, „п”, „т”...). Променом облика вокалног тракта мења се фреквенцијски садржај побуде и као резултат се добијају различити фонеми. Енглески језик разликује 42 фонема.

Проблем говорне анализе представља одређивање параметара екситације и параметара имплусног одзива вокалног тракта. Овај проблем се може назвати и проблемом раздвајања конволуционих компоненти, што је познато под називом деконволуција.

1.1 Кепструм

Кепструм дискретног сигнала $s[n]$ се може израчунати помоћу формуле [2]

$$c_s[n] = \boldsymbol{\mathcal{F}^{-1}}(\log(|\boldsymbol{\mathcal{F}}(s[n])|) \tag{1.3}$$

где су са $\boldsymbol{\mathcal{F}}$ и $\boldsymbol{\mathcal{F}^{-1}}$ означене дискретна Фуријеова трансформација (ДФТ) и инверзна ДФТ. Применом ДФТ на једначину 1.1 добија се

$$S(f) = E(f) \cdot \Theta(f) \tag{1.4}$$

Израчунавање кепструма може се сматрати системом за деконволуцију због чињенице да логаритам производа две компонетне представља збир логаритмованих компоненти

$$\log|S(f)| = \log|E(f) \cdot \Theta(f)| \tag{1.5}$$
$$= \log|E(f)| + \log|\Theta(f)| \tag{1.6}$$
$$= C_E(f) + C_\Theta(f) \tag{1.7}$$

Применом инверзне Фуријеове трансформације добија се кепструм говорног сигнала

$$c_s[n] = \boldsymbol{\mathcal{F}^{-1}}(C_E(f) + C_\Theta(f)) \tag{1.8}$$
$$= \boldsymbol{\mathcal{F}^{-1}}(C_E(f)) + \boldsymbol{\mathcal{F}^{-1}}(C_\Theta(f)) \tag{1.9}$$
$$= c_e[n] + c_\theta[n] \tag{1.10}$$

У облику сигнала $c_s[n]$ уочљиве су области у којима доминирају еквиваленти ваздушне побуде $c_e[n]$, односно импулсног одзива говорних органа $c_\theta[n]$. Утицај ваздушне побуде доминантнији је при већим вредностима аргумента, док ниже вредности аргумента носе информацију о импулсном одзиву вокалног тракта. Због тога се најчешће користи првих 12 кепстралних коефицијената, док нулти коефицијент носи информацију о енергији сигнала.
Пошто је говор реалан сигнал, амплитуда спектра $|S(f)|$ је парна функција и кепструм ће имати реалне вредности

$$c_s[n] = \boldsymbol{\mathcal{F}^{-1}}(\log(|\boldsymbol{\mathcal{F}}(s[n])|) \tag{1.11}$$
$$= \frac{1}{N}\sum_{k=1}^{N-1} \log(|\boldsymbol{\mathcal{F}}(s[n])|) e^{\frac{j2kn\pi}{N}} \tag{1.12}$$
$$= \frac{1}{N}\sum_{k=1}^{N-1} \log(|\boldsymbol{\mathcal{F}}(s[n])|)(\cos\left(\frac{2kn\pi}{N}\right) + j\sin\left(\frac{2kn\pi}{N}\right)) \tag{1.13}$$
$$= \frac{1}{N}\sum_{k=1}^{N-1} \log(|\boldsymbol{\mathcal{F}}(s[n])|)\cos\left(\frac{2kn\pi}{N}\right) \tag{1.14}$$

Због овог својства, често се уместо инверзне дискретне Фуријеове трансоформације користи дискретна косинусна трансформација ради смањења комплексности израчунавања.

1.2 Мел-фреквенцијски кепстрални коефицијенти

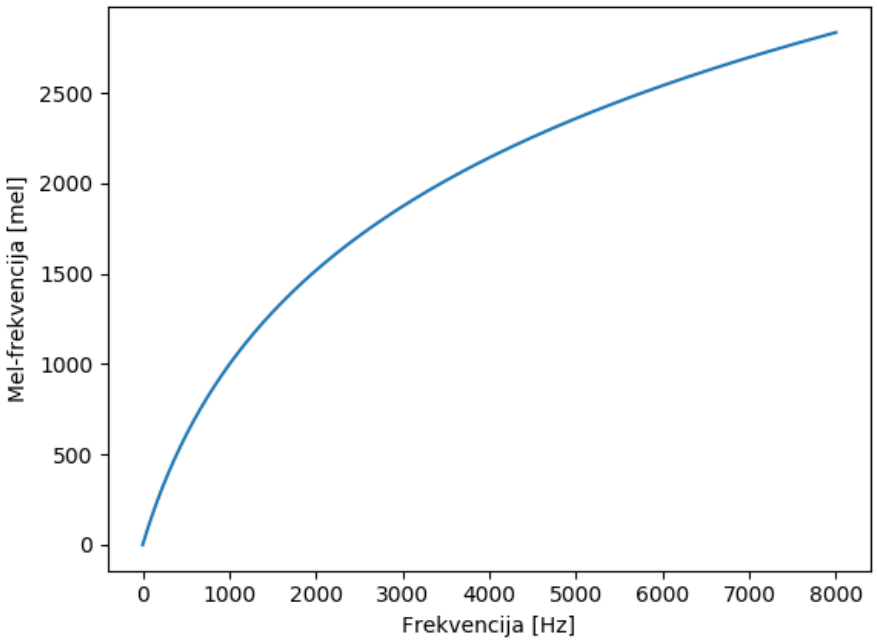
Претходно описани поступци се често користе у применама везаним за аутоматско препознавање говора. У циљу опонашања људског начина доживљавања различитих учестаности у фонетима и примењујући кепстралну анализу настају Мел-фреквенцијски кепстрални коефицијнти (МФКК).

1.2.1 Мелова фреквенцијска скала

Мел-скала је усклађена са људским осећајем висине гласа односно његове учестаности. Њено добијање се врши експериментално, слушаоцу се репродукује тон учестаности 1000 Hz и као његово запажање о висини овог тона се бележи вредност 1000 mel, и ова вредност се користи као мера упоређивања за даље добијање мел скале. Затим се учестаност повећава све док слушалац не примети тон који слуша има дупло већу висину од упоредне вредности и та висина означава вредношћу од 2000 mel. Овај се принцип добија за добијање осталих вредности скале. Експерименти су ипак показали да је међусобна зависност мела и херца линеарна до 500 Hz, док изнад ове учестаности једнаким променама мела одговара све већа промена у херцима. За конвертовање фреквенције у мелову скалу и назад могу се користити формуле

$$M(f) = 1125 \log\left(1 + \frac{f}{700}\right) \tag{1.15}$$

$$M^{-1}(m) = 700\left(\exp\left(\frac{m}{1125}\right) - 1\right) \tag{1.16}$$



Слика 1.1: Мел-фреквенцијска скала

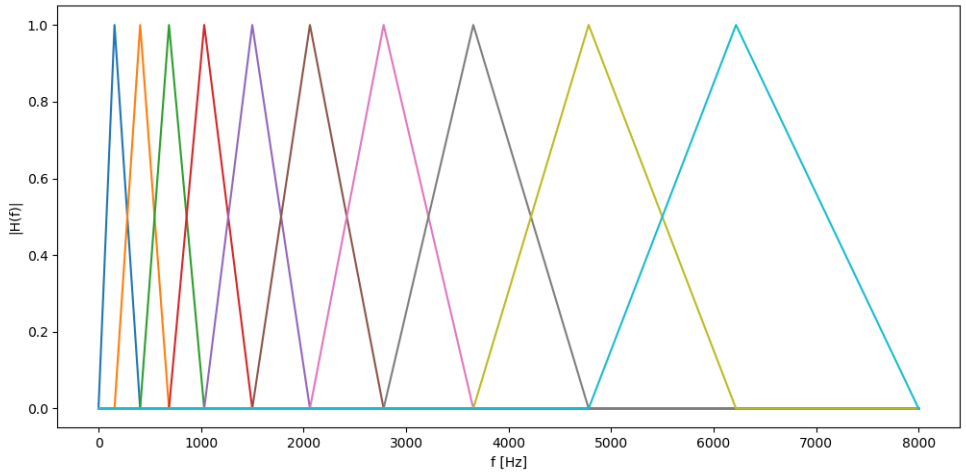
1.2.2 Мелова банка филтара

Постојање аудиторних критичних опсега је такође особина која условљава људски доживљај различитих учестаности. Ова појава везана је за чујни доживљај слушаоца приликом слушања два различита тона на различитим учестаностима. Дакле, уколико се слушоцу пусти тон учестаности f_1 која се налази у неком чујном опсегу, тада слушалац има доживљај тона у складу са мел склаом. Уколико се поред тог тона пусти други тон учестаности f_2 тада звучни доживљај слушаоца зависи од међусобне фреквентне блискости ова два тона. Наиме, уколико је фреквентни размак довољно мали тако да се налазе унутар истог критичног чујног опсега долази до појаве маскирања и слушалац чује тон f_1, али веће гласности. Уколико је фреквентни размак ових тонова већи од ширине чујног критичног опсега тада слушалац чује два тона на одговарајућим претходно поменутим учестаностима. Имајући то у виду фреквенцијски опсег говорног сигнала потребно је изделити на опсеге. Ти опсези формирају мелову банку филтара. Поступак формитања мелове банке филтара се може изделити на кораке:

1. Користећи јендачину 1.15 конвертовати доњу и горњу границу говорног сигнала у мелову скалу. За доњу границу се може узети вредност од 0 Hz, док је горња фреквенција обично условљена Никвистовим креитеријумом. У коришћеном скупу података, говрни сигнал је одабиран са 16 kHz, па је за горњу границу коришћена вредност од 8 kHz.

2. Изделити опсег говорног сигнала на меловој скали на еквидистантне делове (опсеге). За број делова се обично узима број из интервала $[26,\ 40]$. За n филтара добијају се $n+2$ фреквенције на меловој скали.

3. Добијене мелове фреквенције, потребно је вратити у Hz фреквенције, формулом 1.16, које се касније користе за централне учестаности троугаоних филтара



Слика 1.2: Говорни фреквенцијски опсег подељен на 10 мел филтара

1.2.3 Алгоритам израчунавања МФКК

У овом поглављу ће бити приказан детаљан поступак добијања МФКК.

- **Високофреквентно филтрирање**
  Говорни сигнал је по природи аналогни сигнал, који је потребно дискретизовати и дигитализовати да би се израчунала тражена обележја. Поменути процеси су нискофреквентни и утичу на слабљење виших спектралних компоненти у говорног сигналу. Из тог разлога након извршене дигитализације, а пре самог издвајања обележја, потребно је извршити предобраду снимљеног говорног сигнала. То се постиже применом високопропусног филтра првог реда
  $$H(z) = 1 - az^{-1} \tag{1.17}$$
  при чему се параметар a бира из интервала $[0.95,\ 0.98]$[2].

- **Прозоровање сигнала**
  На почетку поглавља је речено да се само кратки сегменти говора могу сматрати као излаз линеарног, временски инваријантног система. Са тим у вези сигнал је потребно изделити на прозоре дужине $20\,\text{ms} - 40\,\text{ms}$ [1]. У овом раду изабрана је дужина прозора од 25 ms, са преклапањем између суседних прозора од 15 ms. Наредни кораци се примењују за сваки прозор.

- **Израчунавање спектра снаге**
  За сваки прозор говорног сигнала је првобитно потребно одредити ДФТ. Ради потискивања бочних лобова, згодно је применити Хамингов прозор пре израчунавања ДФТ. У овом раду ДФТ се израчунава у 512 тачака. Спектар снаге се може естимирати периодограмом, који за сигнал $s[n]$ се израчунава по формули
  $$S(f) = \boldsymbol{\mathcal{F}}(s[n]) \tag{1.18}$$
  $$P(f) = \frac{1}{N}|S(f)|^2 \tag{1.19}$$

- **Филтрирање спектра снаге меловом банком филтара**
  Добијени периодограм је потребно филтрирати кроз сваки филтар из мелове банке. Добијени коефицијенти за сваки филтар се сумирају. На крају се добијају коефицијенти који нам говоре колико је енергије садржано у сваком филтру филтер банке.

- **Логаритмовање**
  Логаритмовати сваки коефицијент добијен пропуштањем периодограма кроз банку филтара.

- **Дискретна косинусна трансформација**
  Дискретном косинусном трансформацијом над логаритмованим енергијама добијају се кепстрални коефицијенти.

1.2.4 Динамичка обележја

Често је од интереса посматрати и промену МФКК у времену [2]. На овај начин, коришћена обележја директно носе информацију о променама између сусед-

них прозора говорног сигнала. Ради израчунавања ових обележја користе се полиномијалне апроксимације првог и другог извода кепстралних коефицијената [3]

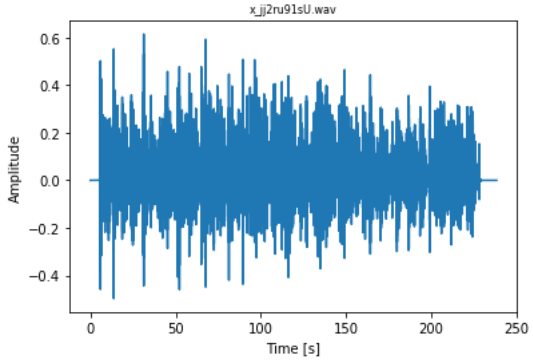
$$\Delta c_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{1.20}$$

$$\Delta^2 c_t = \frac{\sum_{n=1}^{N} n(\Delta_{t+n} - \Delta_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{1.21}$$

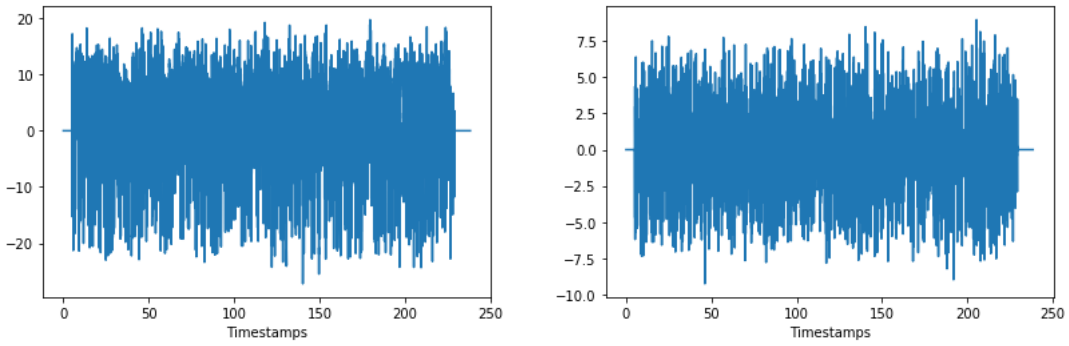
где су Δc_t и $\Delta^2 c_t$ делта и делта-делта кепстрални коефицијенти за прозор t, израчунати из статичких кепстралних коефицијената из околних прозора. Типична вредност која се узима је $N = 2$.

1.3 Приказ обележја говора

У овом раду, као вектор улазних параметара, коришћено је 13 кепстралних коефицијената, са њихом делтама, што даје улазни вектор димензије 26. Уместо првог кепстра, коришћена је енергија сигнала. Из базе видео фајлова, само аудио снимци су извучени уз помоћ алата *FFmpeg* [4], чиме се добија база аудио снимака говора председника Обаме. Након тога, сваки видео је нормализован уз помоћ *FFmpeg – normalize* додатка [5]. Приказ неких кепстралних коефицијената, за насумично одабран аудио снимак из базе, се може погледати у наставку.

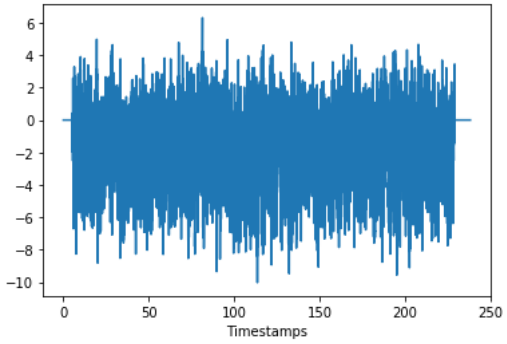


Слика 1.3: Насумични норамлизовани аудио сигнал из базе

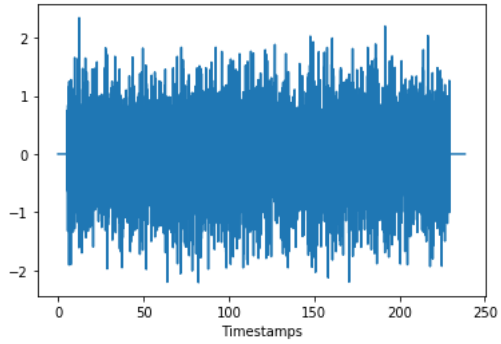


(а) Први МФКК (б) Делта првог МФКК

Слика 1.4: Приказ гласовних обележја

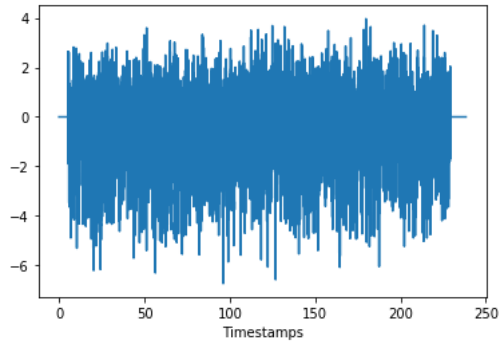


(а) Шести МФКК

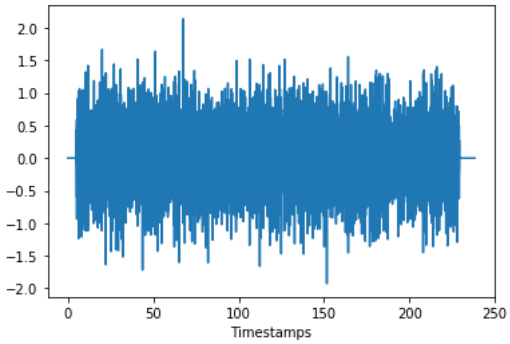


(б) Делта шестог МФКК

Слика 1.5: Приказ гласовних обележја



(а) Дванаести МФКК



(б) Делта дванаестог МФКК

Слика 1.6: Приказ гласовних обележја

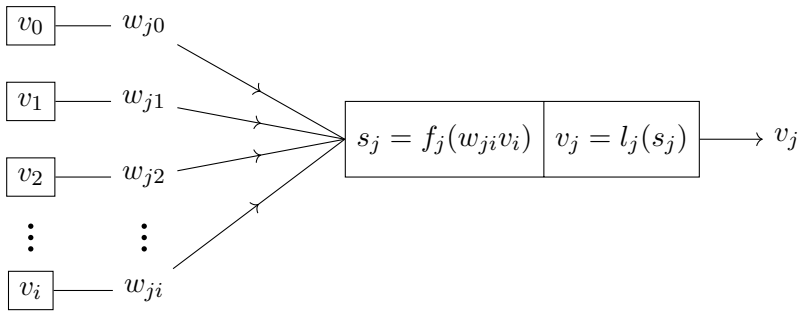
2 | Обележја видеа

3 | Рекурентне неуралне мреже

Неуралне мреже су модели учења који остварају огроман успех у широком спектру задатака супервизираног и несупервизираног машинског учења. Класичне (*feedforward*) неуралне мреже се посебно издвајају у проблемима класификације. Упркос њиховој моћи, стандардне неуралне мреже имају ограничења, од којих је назначајније да примери из базе података морају бити међусобно независни. У случају када су подаци зависни у времену и простору ово није прихватљиво. Фрејмови видеа, делови аудио сигнала, речи из реченице представљају примере података где захтев о међусобној независности није задовољен. Рекурентне неуралне мреже превазилазе тај недостатак и користе се у случахевима када се подаци могу приказати у форми секвенце. Оне имају особину да селективно прослеђују информацију између корака секвенце, док и даље процесују један по један корак. У овом поглављу ће се прво бацити кратак осврт на класичне (*feedforward*) неуралне мреже, а затим ће бити детаљније обрађен појам различитих врста рекурентних неуралних мрежа.

3.1 Преглед вештачких неуралних мрежа

Неуралне мреже су модели израчунавања инспирасни биолошким неуралним системом. Неурална мрежа се састоји од скупа вештачких неурона, који се још и називају чворовима, који су међусобно повезани синапсама (везама). За сваки неурон j се дефинише функција активације $l_j(*)$ и интеграциона функција $f_j(*)$. Веза од чвора j' ка чвору j је описана тежинским коефицијентом $w_{jj'}$.

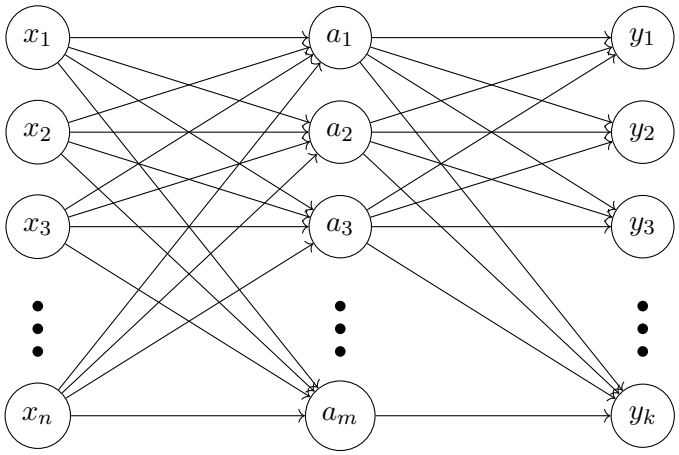


Слика 3.1: Шематски приказ вештачког неурона

За интеграциону функцију се најчешће узима линеарна сума, те се вредност на излазу једног чвора може израчунати као

$$v_j = l_j\left(\sum_{k=0}^{i} w_{ji}v_i\right) \tag{3.1}$$

Чести избори активационе фунцкије су сигмоид $\sigma(z) = 1/(1 + e^{-z})$ и хиперболички тангенс $\tanh(z) = (e^z - e^{-z})/(e^z - e^{-z})$. Од скора, у моделима дубоких неуралних мрежа користи се активациона функција $ReLU(z) = max(0, z)$ која је показала значајно побољшање перформанси.
Feedforward неурлне мреже су врста вештачких неуралних мрежа чији усмерени графови не смеју да садрже петље. Због одсуства петљи, чворове мреже је могуће организовати у слојеве. Излаз једног слоја се добија на основу излаза нижих слојева.



Слика 3.2: Шемтаски приказ вишеслојне *feedforward* неуралне мреже

Вектор обележја $\mathbf{x}$ се доводи на најнижи (улазни) слој. Излази чворова у наредним слојевима се сукцесивно израчунавају, док се не добије излаз највишег (излазног) слоја мреже $\hat{\mathbf{y}}$. *Feedforward* мреже се користе у супервизованом учењу на задацима класификације и регресије. Учење се остварује ажурирањем тежинских коефицијената $w_{jj'}$ са циљем да се оптимизује критеријумска функција $\mathfrak{L}(\hat{\mathbf{y}}, \mathbf{y})$, која пенализује дистанцу између излазног вектора $\hat{\mathbf{y}}$ и вектора циља $\mathbf{y}$.
Најуспешнији алгоритам за транирање неуралних мрежа је алгоритам бекпропагације [6]. Алгоритам бекпропагације користи ланчано правило за рачунање извода критеријумске функције $\mathfrak{L}$. Тежински коефицијенти се поправљају користећи алгоритам градијентног спуста. Најчешће коришћен алгоритам градијентног спуста је градијентни спуст који користи мини чаршже (енг. *batch*) обучавајућег скупа. Тај алгоритам комбинује предности шаржног и стохастичког градијентног спуста. Применом тог алгоритма коефицијенти се ажурирају на основу акумулиране грешке свих примерака из мини шарже. За шаржу дужине n правило ажурирања тежинских коефицијената се може записати овако

$$w := w - \eta \nabla_w \mathfrak{L}\left(\hat{\mathbf{y}}^{\mathbf{i:i+n}}, \mathbf{y}^{\mathbf{i:i+n}}\right) \tag{3.2}$$

где је са η означена брзина обучавања (енг. *learning rate*). У опстем случају, критеријумска функција није конвексна, и не постоји гарант да ће градијентни спуст довести до глобалног минимума. Многе варијанте градијентног спуста се уводе како би се убрзало обучавање. Неке од најпопуларнијих су: *AdaDelta* [7], *AdaGrad*

[8], *RMSprop* [9] и *Adam* [10]. Углавном се заснивају на мењању брзине обучавања као и на моментуму који представља промену тежинског коефицијента у претходној итерацији алгоритма.
Да би се израчунао градијент у једначини 3.2 користи се, већ поменут, алгоритам бекпропагације. Као што јој само име каже, алгоритам започиње од излазног слоја и креће уназад ка нижим слојевима. За тежинске коефицијенте за везе између излазног и скривеног слоја може се писати

$$\Delta w_{oh} = -\eta \frac{\partial \mathfrak{L}\left(\hat{\mathbf{y}}, \mathbf{y}\right)}{\partial w_{oh}} \tag{3.3}$$

$$= -\eta \left[\frac{\partial \mathfrak{L}\left(\hat{\mathbf{y}}, \mathbf{y}\right)}{\partial \hat{\mathbf{y}}}\right] \left[\frac{\partial \hat{\mathbf{y}}}{\partial s_j}\right] \left[\frac{\partial s_j}{\partial w_{oh}}\right] \tag{3.4}$$

$$= \eta \delta_{oh} v_h \tag{3.5}$$

За тежински коефицијент за везе између два излазна слоја може се питати

$$\Delta w_{oh} = -\eta \frac{\partial \mathfrak{L}\left(\hat{\mathbf{y}}, \mathbf{y}\right)}{\partial w_{oh}} \tag{3.6}$$

$$= -\eta \left[\frac{\partial \mathfrak{L}\left(\hat{\mathbf{y}}, \mathbf{y}\right)}{\partial \hat{\mathbf{y}}}\right] \left[\frac{\partial \hat{\mathbf{y}}}{\partial s_j}\right] \left[\frac{\partial s_j}{\partial w_{oh}}\right] \tag{3.7}$$

$$= \eta \delta_{oh} v_h \tag{3.8}$$

4 | Обучавање модела

5 | Синтеза видеа

Литература

[1] Ђуровић, Ж. Материјали са предмета Обрада и препознавање говора

[2] Делић, В. Анализа мел-фреквенцијских кепстралних коефицијената као обележја коришћених при аутоматском препознавању говорника

[3] http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[4] https://ffmpeg.org/

[5] https://github.com/slhck/ffmpeg-normalize

[6] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning internal representations by error propagation. Technical report, DTIC Document, 1985.

[7] Matthew D. Zeiler. Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701, 2012.

[8] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. The Journal of Machine Learning Research, 12:2121–2159, 2011.

[9] Tijmen Tieleman and Geoffrey E. Hinton. Lecture 6.5- RMSprop: Divide the gradient by a running average of its recent magnitude, 2012.

[10] Diederik P. Kingma, Jimmy Ba, Adam: A Method for Stochastic Optimization, 2014.