

KeySniff: Gépelt szöveg feltárása billentyűzet hangfelvételekből mély tanulással

vajdad4m

Vajda Ádám

Felkészítő tanár: Dr. Vajda Róbert

ELTE Radnóti Miklós Gimnázium, 1146 Budapest, Cházár András u. 10.

1. Bevezetés

1.1. Célkitűzés

A projekt célja egy „KeySniff” nevű innovatív megoldás kifejlesztése, amely gépi tanulási és mesterséges intelligencia technikákat alkalmaz a gépelt szöveg pontos rekonstruálására a billentyűzethangok hangfelvételeiből. A kutatás célja, hogy megoldja a gépelt szöveg rekonstruálásának kihívását a billentyűzethangokról készült hangfelvételekből, amelyek gyakran zajosak.

1.2. Téma aktualitása

A gépelt szöveg rekonstruálásának gondolata a billentyűzetről készült hangfelvételekből rendkívül fontos és jelentős a mai digitális korban. Ez a technika billentyűzeteken kívül alkalmazható ATM terminálokon, vagy esetleg gombos beléptető rendszereken.

Fontos azonban megjegyezni, hogy ehhez a technológiához jelentős biztonsági kockázat is kapcsolódik. Ha az emberek képesek a szöveget, például jelszavakat, személyes információkat vagy érzékeny adatokat csak a gépelés hallatán visszanyerni, az komoly veszélyt jelenthet a magánéletre és a biztonságra. Ezért ennek a technológiának az alkalmazhatóságát alaposan meg kell fontolni, és megfelelő biztonsági intézkedéseket kell bevezetni a felhasználói adatok védelme érdekében.

1.3. Kapcsolódó munkák

Már készültek projektek billentyűzetek akusztikus lehallgatására. Először 2004-ben tettek próbálkozást (Asonov & Agrawal, 2004) számítógépek billentyűinek, nyomógombos telefonok és ATM gombok lehallgatására. FFT-vel alakítják át a hangot a modellbe bemenő jellemzőkre. Az adathalmaz méretét nem említették. 2009-ben publikáltak egy újabb kutatást (Zhuang, Zhou, & Tygar, 2009) az akusztikus lehallgatás megvalósítására. 12 percnyi hangfelvétel kellett ahhoz, hogy a modell elérjen 90% feletti pontosságot. Egy ma fejlesztett nyílt forráskódú megoldás a Keytap3 (Gerganov, 2022). Ez mélytanulás helyett egy statisztikai modellt állít fel a lenyomott billentyűk összehasonlítására és kiválasztja a leghasonlóbbat.

WiFi jelek alapján is képesek voltak 2015-ben feltanítani egy modellt, ami képes meghatározni a lenyomott billentyűt. A publikáció (Ali, Liu, Wang, &

Shahzad, 2015) betűkként 30 mintáról számol be, ami 26 billentyű esetén 780 mintának felel meg. Ennek felvétele körülbelül 6 percet vesz igénybe. Ezen az adathalmazon a módszer 83,46% pontosságot ér el.

1.4. Kutatási terv

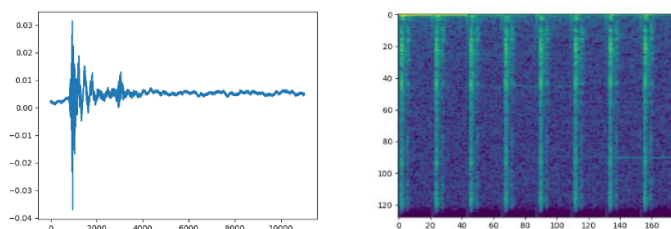
A „KeySniff” rendszer megvalósításának első lépése a billentyűzethangok hangfelvételeinek nagy adathalmazának összegyűjtése. Az adathalmaz gyűjtésén egyszerre történik a mikrofon folyamatos rögzítése, illetve a lenyomott billentyűk karakterének és a lenyomás pillanatának rögzítése. Ezután az összegyűjtött adatokon automatikusan elvégezzük a hangfelvételen a szükséges feldolgozást. Ez tartalmaz vágását, majd spektrogrammá alakítást.

A következő lépés a mély tanulási algoritmusok betanítása. A tanítási folyamat során az algoritmusok rendelkezésére bocsátjuk a spektrogram képeket és a modell kimeneti céljaként az adott pillanatban lenyomott billentyűt várjuk.

2. A kutatási folyamat

2.1. Adatgyűjtés és előfeldolgozás

A gyűjtési folyamatban egyidejűleg rögzítünk hangot egy mikrofon által a Pyaudio könyvtárral és a lenyomott billentyűket a Pygame könyvtárral. A lenyomott billentyűk mind az angol ábécé valamelyike, illetve a szóköz karakter. Így 27 lehetséges karakterből álló betűkeretet kapunk. Az adathalmaz billentyűlenyomásokként 0,5 másodperces részletek. A részletek a lenyomás pillanatától 0,1 másodperccel hamarabb kezdődnek, illetve 0,4 másodpercig tartanak. Ezeket a részleteket Mel Spektrogrammá alakítjuk. Így 1024 x 22-es felbontású képeket kapunk, amiket át tudunk alakítani 128 x 176 pixel méretű képekre.



1. ábra: Billentyűlenyomás és hozzá tartozó átméretezett spektrogram

Összesen 600 mintát gyűjtöttem, ami 5 percnyi hangfelvételnél felel meg, amik közel azonos billentyűzet-mikrofon helyzetből készültek. Egy általam készített segédprogram által tudtam figyelni a lenyomott billentyűk egyenletes elosztására

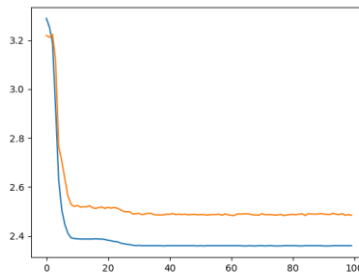
2.2. Modellezés

Az előállított modell bemenete egy 128 x 176-os méretű 1 színcsatornás spektrogram kép, a kimenete, pedig egy 27 nagyságú vektor, ami az egyes billentyűnek a valószínűsége. Mivel a modell célja, hogy valószínűségeket becsüljön a végső réteg aktiválási funkciója a softmax. A modell veszteségfüggvénye, pedig a kategorikus kereszt-entrópia. Architektúrában a VGG-16 (Simonyan & Zisserman, 2014)-ra van alapozva. Ez egy képosztályozásra készített modell szerkezet. Mivel a hangot képpé alakítottuk, kezelhetjük a feladatot képfeldolgozásként, így a konvolúciós neurális hálók (CNN-ek) alkalmazása lehetőséget biztosít a probléma megoldására.

A modell túlillesztésének elkerülése érdekében több technikát alkalmaztam. Egyik a tanító bemeneti adatok véletlenszerű torzítása, azaz eltolása és zaj hozzáadása. Másik a 0,2-es valószínűségű Dropout rétegek használata. A modell tanítása során lineáris tanulási ráta csökkentést is alkalmaztam, illetve 80-20-as tanítás-ellenőrzési eloszlást, ahhoz, hogy megfigyelhető legyen a modell teljesítménye ismeretlen adatpontokon.

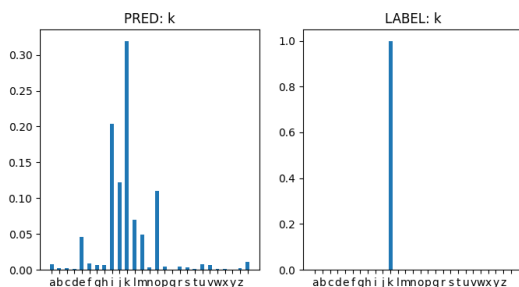
3. Eredmények

A modellt 16-os kötegmérettel tanítottam egy NVIDIA RTX 3060 videókártyán. A modell a tanulás során elkerülte a túlillesztést. Végül 100 tanulási lépés után 96,4 %-os pontosságot ért el a validációs adathalmazon.



2. ábra: A modell tanulási görbéje

A tanulási görbén megfigyelhető, hogy a modell nagyon hamar konvergál a tanítási adat globális minimumára, így a validációs veszteség azon pont után nem képes csökkenni. Ennek kiküszöbölésére lehet több adatot gyűjteni. Ez azonban szembe menne a projekt céljával, ami pedig, az, hogy kevés adaton (5 percnyi hangfelvételen) is képesek legyünk elfogadható pontossággal (>90%) megbecsülni a lenyomott billentyűket.



3. ábra: A modell egy becslése

A modellnél megfigyelhető, hogy valóban a helyet próbálja becsülni, mert a többi rosszul becsült karakter a valós cél, azaz a „k” karakter közelében helyezkednek el a billentyűzetten.

4. Jövőbeli fejlesztések

Lehetséges és ígéretes fejlesztésnek tartom a kamera alapú távolság- és mélységbecslés területén jól bevált sztereókamerás, a mi esetünkben sztereó mikrofonos adatok használatát. Ez hasonlóképpen utánozná az emberi biológiát, mint a látásnál a sztereó szenzorjaink, azaz a két szemünk. A két fül segíti az embereket a hangok irányának meghatározására, mivel képesek vagyunk észlelni a kettő érzékelés közötti eltérést, így tudunk becsülni irányt. Másik fejlesztés, pedig a címkézetlen adatokon felügyeletlen tanítás alkalmazása. Ilyenkor a modell két bemeneti spektrogramot hasonlít össze és kimenetül egy eltérési vektort ad, ami a billentyűzetten a két billentyű relatív eltérése.

5. Összefoglalás

Kutatásom célja, **hogy felhívjam a figyelmet az akusztikus lehallgatás potenciális veszélyére.** Nem kizárólag billentyűzetten, hanem más digitális eszközökön is veszélyforrást jelent. Bemutattam egy módszert, amely képes megjósolni a lenyomott billentyűket nagy pontossággal (96,4%). A megközelítésem kis mennyiségű gyűjtött adatot használ, mindössze 5 percnyi hangfelvételt. Ismertettem az általam használt módszer lehetséges jövőbeli fejlesztéseit, ami még nagyobb pontosságot eredményez.