Mesterséges Intelligencia Kutatócsoport
Szegedi Tudományegetem

Arenberg Doctoral School
Faculty of Engineering Science
KU Leuven

Spektro-temporális feldolgozási módszereken alapuló zajtűrő automatikus beszédfelismerés

A PhD-értekezés tézisei

Kovács György

Témavezetők:

Dr. Tóth László
Prof. dr. ir. Dirk Van Compernolle

Szeged
2017

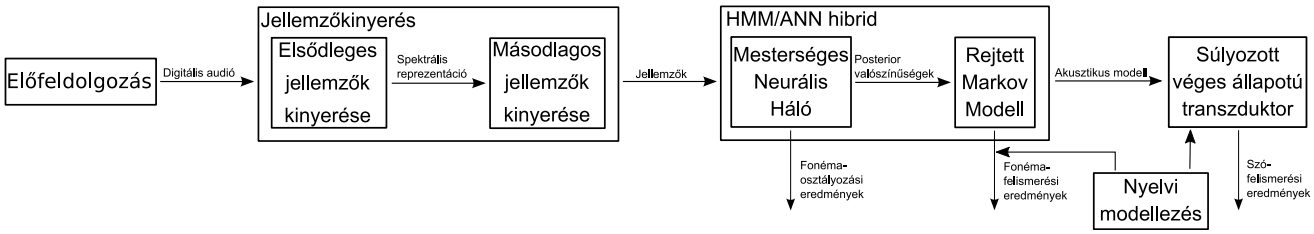
1. Bevezetés

Automatikus beszédfelismerés (ASR) során azt várjuk a géptől, hogy a beszédjelet automatikusan szavak vagy fonémák sorozatává írja át. Ennek a látszólag egyszerű problémának számos felhasználási területe van a diktálórendszerektől kezdve, a dialógus rendszereken át, egészen a személyi asszisztens alkalmazásokig. Ezen alkalmazások felhasználhatósága számos kritériumtól függ, úgy mint a sebesség, valamint a számítási- és memóriaigény. Jelen munkánk középpontjában azonban a pontosság áll, különös tekintettel a zajjal szennyezett és eltérő átviteli karakterisztikával felvett beszédre. Az a tény teszi különösen nagy kihívássá a zajtűrési kritériumnak való megfelelést, hogy gyakran lehetetlen előre látni az összes potenciálisan felmerülő nehézséget. Így a beszédfelismerő rendszereket úgy kell megtervezni, hogy olyan körülmények között is működjenek, mely körülményekre nem lettek külön felkészítve.

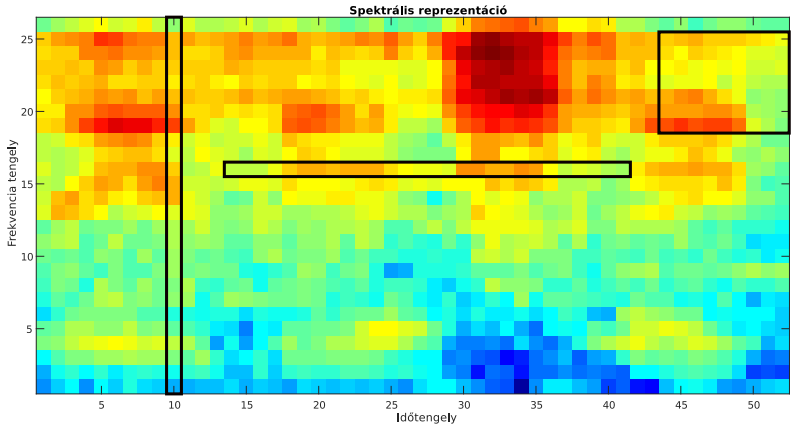
Jelen munkában ezt a problémát a HMM/ANN modell keretében vizsgáljuk, melynek főbb lépéseit mutatja be az 1. ábra. Ez a megközelítés abban különbözik a hagyományos HMM/GMM modelltől, hogy a generatív GMM (Gaussian Mixture Model) modellt diszkriminatív neuronhálómodell váltja. A HMM/ANN hibrid modell évtizedek óta jelen van [33], de alkalmazása napjainkban lett különösen népszerű, a mély neuronhálók megjelenésének köszönhetően. Az alap HMM/ANN modelltől jelen munkában abban térünk el, hogy – követve a Kleinschmidt használta elnevezéseket [17] – a jellemzőkinyerés folyamatát elsődleges- és másodlagos jellemzőkinyerésre bontjuk. Ezt a felosztást egyrészt azért eszközöltük, mert míg az elsődleges jellemzőkinyerés során korábban ismert, jól bevált módszerek használatára szorítkoztunk, a másodlagos jellemzőkinyerés folyamatában több módosítást is javasoltunk. Az akusztikus modell mellett a másodlagos jellemzőkinyerés lépése volt a folyamat azon része, melyre koncentráltunk ebben a munkában, akár azzal, hogy módosításokat vezettünk be, akár azzal, hogy azt összevontuk az akusztikus modell tanításával.

Módosítási javaslataink mind a felismerés pontosságának növelését célozták, különös tekintettel a zajos környezetben történő beszédfelismerésre. Egyes kutatók ezt a célt az emberi beszédértésről szerzett ismeretek felhasználásával próbálják elérni. Ezt a megközelítést az az elképzelés motiválta, hogy ha tanulmányozzuk a jobban teljesítő emberi beszédpercepciót [41], az így tanultakkal bizonyára javítani tudunk a rosszabbul teljesítő automatikus beszédfelismerésen is. Bizonyos fokig ennek a megközelítésnek a sikerét láthattuk például a Mel-skála vagy a Bark-skála alkalmazásánál, ahogy a spektro-temporális módszerek alkalmazásában is, melyre az itt bemutatott módszerek támaszkodnak.

A javasolt módszereket három feladaton értékeltük ki: fonémaosztályozás, fonémafelismerés, és szófelismerés. Ehhez a Szeged magyar nyelvű híradós adatbázist [46], valamint a TIMIT [28] és az Aurora-4 [36] angol nyelvű beszédadatbázisokat használtuk fel.



1. ábra. A beszédfeldolgozási folyamat áttekintése.



2. ábra. Az Aurora-4 adatbázisból vett példa spektrális reprezentációja. A fekete keretek (balról jobbra) jelzik az a) MFCC b) TRAP és c) lokalizált spektro-temporális jellemzők alakját.

2. Spektro-temporális jellemzőkinyerés

Miközben egyre többet tudunk az emberi beszédértésről, a hagyományosan használt jellemzőkinyerési módszerek továbbra is csak annak legalapvetőbb tulajdonságait veszik figyelembe. Ésszerű elvárás azonban az emberi beszédpercepció jobb teljesítményét figyelembe véve, hogy egy annak tulajdonságait jobban közelítő rendszer jobban teljesítsen, mint egy pusztán matematikai eszközökre épülő módszer. Egy ilyen tulajdonság, melynek közelítéséből profitálhat az automatikus beszédfelismerés, a kortikális sejtek együttes spektro-temporális érzékenysége [6].

Annak ellenére, hogy az emberek alig képesek felismerni ilyen rövid részleteket, az MFCC jellemzőkinyerés 20-30 ezredmásodperces részletekben dolgozza fel a beszédjelet. Ugyan ez a környezet kiterjeszthető a Δ együtthatók és szomszédos keretek (frame) felhasználásával, ez mégis különbözik attól, mint ha a jellemzőket egy adott spektro-temporális modulációra hangolnánk [4]. Ráadásul a felhasznált keretek számának növekedésével a jellemzők száma is nő, ami dimenziócsökkentő módszerek alkalmazását teheti szükségessé. További probléma az MFCC használatával, hogy az eredményül kapott jellemzővektorok globálisak a frekvenciatartományban, így sávhatárolt zaj esetén is minden jellemző sérül [4]. Egyes kísérletek szerint az emberi beszédértés relatív szűk frekvenciasávokon alapul [32], ami sugallja, hogy az ablakoknak (melyekből a jellemzőket kinyerjük) a frekvenciatartományban is korlátozottnak kell lenniük. Ugyanez a megfontolás motiválta a TRAP modell bevezetését [12], ahol minden frekvenciasáv külön kerül feldolgozásra. A 2. ábrán szemléltetjük ezt a megközelítést, a hagyományos MFCC megközelítés mellett. Ezen megközelítésekkel szemben jelen tanulmányban a spektro-temporális feldolgozásra koncentrálunk, ahol a spektrális reprezentációt olyan ablakokban dolgozzuk fel, melyek az idő- és frekvenciatartományban is behatároltak (lásd 2.(c) ábra).

A spektro-temporális feldolgozás folyamatát tekinthetjük úgy, mint F szűrő alkalmazását P ablakra, ahol az o kimenet a következő formula segítségével kapható meg:

$$o = \sum_{f=0}^{M-1} \sum_{t=0}^{N-1} P(f,t)F(f,t), \tag{1}$$

ahol M és N adja az F szűrő és P ablak magasságát és szélességét. A formula alkalmazásával könnyen kinyerhetjük jellemzők egy csoportját különböző szűrők használatával, vagy ugyanazt a szűrőt más-más ablakra alkalmazva. Ám nehéz megtalálni azokat a szűrőparamétereket, melyek az adott feladat esetén optimális eredményt adnak.

2.1. 2D DCT

Az MFCC feldolgozás lokalizált spektro-temporális ablakokra történő általánosításának egy kézenfekvő módja a feldolgozás során alkalmazott diszkrét koszinusz-transzformáció (DCT) helyettesítése annak kétdimenziós változatával (2D DCT). Az így keletkező jellemzőkinyerési módszert tekinthetjük úgy mint az (1) egyenlőség alkalmazása az alábbi F szűrő behelyettesítésével:

$$F_{pq}(f,t) = \cos\frac{\pi\cdot(2f+1)\cdot p}{2M}\cos\frac{\pi\cdot(2t+1)\cdot q}{2N}, \tag{2}$$

ahol M és N a szűrő magassága és szélessége; ezt a tartományt járják be az f és a t paraméterek, míg a p és a q paraméterek specifikálják a szűrők modulációs frekvenciáját a frekvencia- és időtartományban.

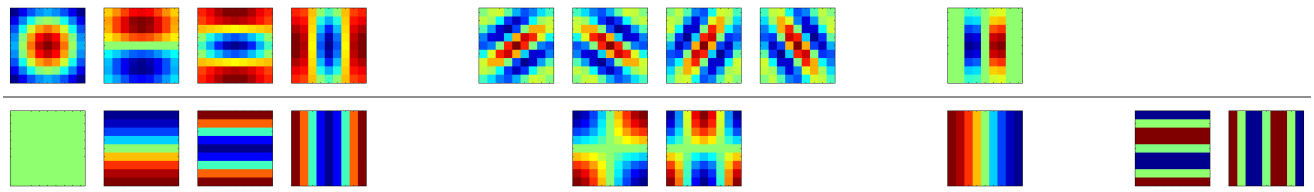
Definíció szerint egy $M \times N$ méretű ablakra a 2D DCT ugyanannyi ($M \cdot N$) együtthatót ad vissza. Több eredmény is (mind a beszédfelismerés, mind a képfeldolgozás területéről) azt jelzi azonban, hogy ezek az együtthatók nem egyformán fontosak az ablak tartalmának reprezentációja szempontjából (az alacsonyrendű együtthatók fontosabbak) [4, 16]. Eldöntetlen kérdés azonban, hogy pontosan hányat lehet érdemes megtartani az alacsonyrendű együtthatók közül. Kísérleti úton próbáltunk választ találni erre a kérdésre. Továbbá kísérleti úton vizsgáltuk az ablakok méretének kérdését, az ablakok átfedésének mértékét, és a felhasználandó mel-szűrők számát is. Ennek érdekében számos fonémafelismerési kísérletet végeztünk a TIMIT adatbázison. Ezen kísérletek eredményeinek felhasználásával kiválasztottunk három lehetséges paraméterbeállítást. Végül (részben az MFCC eredményekkel való jobb összehasonlíthatóságot is figyelembe véve) úgy döntöttünk, hogy a logaritmikus spektrumot használjuk bemenetként 26 mel-szűrővel. Továbbá úgy döntöttünk, hogy a bemeneten használt ablakok 7 csatorna magasak, és 9 keret szélesek lesznek, és a 12 pozíción helyezzük el őket a frekvenciatartományban, minden pozíción 9 együtthatót kinyerve.

A kiválasztott beállításokat a TIMIT beszédadatbázison értékeltük ki a „mag” (core) teszthalmaz tiszta, valamint zajjal szennyezett változatain. A spektro-temporális jellemzőkinyeréssel (2D DCT) elért eredményeket összehasonlítottuk azokkal, melyeket az MFCC jellemzők felhasználásával kaptunk (lásd 1. táblázat). Az eredmények azt mutatják, hogy a spektro-temporális feldolgozással alacsonyabb hibaarányokat kapunk zajos beszéd esetén (a zaj mértékétől függetlenül), mint az MFCC együthatók használatával.

Bemutattunk továbbá egy módszert a két jellemzőkészlet kombinációjára, ahol a két jellemzőkészeletre függetlenül egy-egy neuronhálót tanítunk, majd a neuronhálók kimenetét egy harmadik neuronhálóban használjuk fel. Megmutattuk, hogy a két jellemzőkészlet kombinációjával tiszta beszéd esetén is jobb eredményt tudunk elérni.

Jellemzőkészlet	Tiszta beszéd	Zajjal szennyezett beszéd					
		Rózsaszín zaj			„Babble” zaj		
		20 dB	10 dB	0 dB	20 dB	10 dB	0 dB
MFCC	29.11%	56.78%	74.78%	85.57%	48.04%	73.56%	86.29%
2D DCT	29.52%	46.62%	67.01%	79.07%	41.03%	58.36%	74.81%

1. táblázat. Fonémafelismerési hibaarányok a TIMIT core teszthalmaz tiszta és zajjal szennyezett változatán.



3. ábra. Gábor-szűrők manuálisan összeállított készlete (fenn), és a 2D DCT szűrőkészlet (lenn), a hasonlóság kiemelése érdekében a megfelelő szűrők egymás közelébe pozicionáltuk. A szűrők mérete 9×9. [26]

2.2. Gábor-szűrők

Egy másik módszer a spektro-temporális jellemzők kinyerésére a Gábor-szűrők valós részének alkalmazása az ablakokra, melyet egy két-dimenziós Gauss-görbe,

$$W(f,t) = \frac{1}{2\pi\sigma_f\sigma_t} e^{-\frac{1}{2}\left(\frac{(f-f_0)^2}{\sigma_f^2}+\frac{(t-t_0)^2}{\sigma_t^2}\right)}, \tag{3}$$

és egy orientált szinusz, azaz

$$S_{\Omega,\omega}(f,t) = cos\left(\frac{\pi\cdot f\cdot 2\Omega}{M}+\frac{\pi\cdot t\cdot 2\omega}{N}\right) \tag{4}$$

szorzataként kapunk. Itt az f és a t paraméterek az ablakok frekvencia- és időtartományát járják be. Továbbá az f_0 és a t_0 paraméterek határozzák meg a Gauss-görbe középpontját, míg a σ_f^2 és a σ_t^2 paraméterek a görbe szélességét határozzák meg a megfelelő tartományokban. Végül, M és N határozzák meg a szűrő méreteit, míg az Ω és az ω paraméterek a szinuszoid dőléséért és periodicitásáért felelnek.

Habár az világos, hogy az Ω és az ω paraméterek hogyan határozzák meg a Gábor-szűrők alakját, kevésbé nyilvánvaló, hogyan kéne ezen paramétereknek értéket választani. Ahogy az sem nyilvánvaló, hogy hány és mekkora szűrők alkalmazása szükséges a megfelelő beszédfelismerési eredmény érdekében. Ezen kérdések megválaszolásával többen foglalkoztak az évek során [8, 18, 42], és ezért mi is részletesen vizsgáltuk. Először leírtunk két automatikus jellemzőkiválasztási módszert a feladatra, név szerint a Kleinschmidt és Gelbart által is használt Filter Finding Neural Network (FFNN) [18] és a Pudil és társai által bemutatott Sequential Forward Floating Selection (SFFS) [38] módszert. Továbbá létrehoztunk két szűrőkészletet ezen módszerek segítségével. Később, a TIMIT adatbázison és a Szeged magyar nyelvű híradós adatbázison végzett kísérletek eredményei alapján összehasonlítottuk a következő Gábor-szűrőkészleteket:

- A Kleinschmidt és Gelbart által bevezetett Gábor-szűrőkészletek [18] (G1, G2, G3)
- A Schädler és társai által bevezetett Gábor-szűrőkészlet [42] (SMK)
- Az általunk, az FFNN módszerrel előállított Gábor-szűrőkészlet (FFNN)
- Az általunk, az SFFS módszerrel előállított szűrőkészlet (SFFS)
- Az általunk, egyszerű heurisztikák – mint a 2D DCT együthatókkal mutatkozó hasonlóság (lásd 3. ábra) – alapján, manuálisan előállított Gábor-szűrőkészlet (Gábor manuális)
- Tíz, véletlenszerűen generált Gábor-szűrőkészlet

Az összehasonlítást először a TIMIT adatbázison végeztük el. Azt találtuk, hogy a manuálisan előállított Gábor-szűrőkészlet jobban teljesített minden más vizsgált szűrőkészletnél, ide értve azokat is, melyek kifinomult automatikus jellemzőkiválasztási módszerek eredményeként jöttek létre. A tiszta és zajos beszéden végzett kísérletek is ugyanerre az eredményre vezettek. Úgy találtuk továbbá, hogy sok esetben a véletlenszerűen generált szűrőkészletek hasonlóan vagy akár jobban teljesítettek, mint kifinomult módszerekkel létrehozott társaik. És hasonlóan ahhoz, mint amit a 2D DCT jellemzőkkel végzett kísérletek esetén tapasztaltunk, Gábor-szűrők esetén is azt tapasztaltuk, hogy egy megfelelő spektro-temporális jellemzőkészlettel jobb eredményeket érhetünk el az MFCC jellemzőkkel elért eredményeknél, akár tiszta, akár zajos beszéd esetén.

Kísérleteinket megismételtük a Szeged magyar nyelvű híradós adatbázison. A Gábor-szűrőkészletek egymáshoz viszonyított teljesítményét tekintve hasonló eredményeket kaptunk, mint a TIMIT adatbázis esetén. Egyrészről a manuálisan létrehozott Gábor-szűrőkészlet alacsonyabb hibaarányokat produkált, mint bármely más vizsgált szűrőkészlet. Ez arra utalhat, hogy a szűrőkészlet létrehozásánál használt heurisztikák általánosabbak az egyéb vizsgált jellemzőkiválasztási módszerekben alkalmazottaknál. A különböző jellemzőkiválasztási módszerekkel készült szűrőkészletek eredményeinek összehasonlításánál azt láttuk, hogy az SFFS algoritmus által előállított szűrőkészlet jelentősen túlszárnyalta az FFNN algoritmus által előállított társát. Mivel hasonló tendenciát figyeltünk meg a TIMIT adatbázison végzett kísérleteknél is, az SFFS jellemzőkiválasztási módszer jobbnak tűnik, mint az FFNN. De még az SFFS algoritmussal előállított szűrőkészlet is rosszabbul teljesített, mint az MFCC jellemzők, vagy akár a manuálisan előállított Gábor-szűrőkészlet. Ugyanez mondható el a Kleinschmidt és Gelbart által bemutatott szűrőkészletekről (G1, G2, G3), vagy a Schädler és társai által összeállított szűrőkészletről is. Ez azt mutatja, hogy az automatikus jellemzőkiválasztási módszerek nemcsak időigényesek, de az eredményül kapott szűrőkészlet eltérő adatbázisra/nyelvre történő adaptációja sem nyilvánvaló. Továbbá azt is láttuk, hogy a legjobb véletlenszerűen generált szűrőkészlet hasonlóan teljesített, mint az SFFS készlet, és jobb eredményeket adott, mint a G1, G2, SMK, vagy az FFNN készlet. Ez ismét a jellemzőkiválasztási módszerek kudarcát mutatja.

A fejezet eredményeinek tézisszerű összefoglalása

I/1. Megmutattuk, hogy 2D DCT jellemzőkinyerés a konvencionális mel-szűrős spektrális reprezentáción is elvégezhető oly módon, hogy hasonló vagy jobb eredményeket kapjunk, mint az MFCC jellemzők használatával. Továbbá azt találtuk, hogy az előbbi előnye jobban kimutatható zajos beszéd esetén (publikálva: [19, 20]).

I/2. Bemutattunk egy egyszerű, ám hatékony stratégiát a konvencionális (MFCC) jellemzők, és a spektro-temporális (2D DCT) jellemzők kombinálására. Megmutattuk a TIMIT adatbázis core teszthalmazának tiszta verzióján, hogy a kombináció jobb eredményt ad, mint a jellemzőkészletek külön külön (publikálva: [20]).

I/3. Bemutattunk és kiértékeltünk egy Gábor-szűrőkészletet. Megmutattuk a TIMIT adatbázison, valamint a Szeged magyar nyelvű híradós adatbázison, hogy ez az itt bemutatott szűrőkészlet jobban teljesít, mint a korábban bemutatott Gábor-szűrőkészletek, valamint az összehasonlításul jelen munka keretében létrehozott szűrőkészletek. Megmutattuk továbbá a TIMIT adatbázison, hogy angol nyelvű beszéd esetén a bemutatott szűrőkészlet jobban teljesít – mind tiszta, mind zajjal szennyezett beszéden – mint az MFCC jellemzők, míg magyar nyelvű beszéd esetén az MFCC jellemzők használatával kapott eredményekhez hasonló eredményeket érhetünk el a használatával (publikálva: [26]).

3. A spektro-temporális jellemzők és a neuronháló együttes optimalizálása

A hagyományos megközelítésben az osztályozók tanítása, valamint a jellemzők előre meghatározott készletének kinyerése elkülönül. Spektro-temporális beszédfeldolgozás esetén ez azt jelenti, hogy a másodlagos jellemzők kinyerése és az akusztikus modell tanítása két elkülönülő lépésben történik. Bár a lépések ilyetén való szétválasztása kényelmes, az optimálistól elmaradó jellemzők használatához vezethet.

Könnyen láthatjuk ennek a megközelítésnek a hibáját, amikor a jellemzőkészlet összeállítása manuálisan történik: a felismerés pontossága nagyban függ az emberi szakértő képességétől egy hatékony jellemzőkészlet összeállítására. A Gábor-szűrőkkel kapcsolatos korábbi eredményeink megmutatták a megközelítés hiányosságait automatikusan összeállított jellemzőkészletek esetére is: a jellemzőkiválasztási algoritmusok nemcsak lassúnak bizonyultak, de nem sikerült olyan jellemzőkészletet előállítaniuk, amely több adatbázison is sikeresen alkalmazható lett volna. Sőt, az általunk manuálisan létrehozott szűrőkészlet nemcsak az eltérő adatbázison végzett kísérletekben teljesített jobban, hanem sokszor azon az adatbázison is, melynek felhasználásával az automatikusan létrehozott jellemzőkészlet összeállításra került. Sajnálatos módon a manuális összeállításnál használt heurisztika sem nyújt biztosítékot az eredményül kapott jellemzőkészlet optimalitására.

A fenti okok miatt javasoltuk a jellemzőkiválasztási és a statisztikai modellezés lépéseinek összevonását. Ez hasonló koncepció, mint amit egy újabb keletű tanulmányban láthattunk, ahol a konvolúciós neuronhálót (CNN) kiterjesztették a jellemzőkinyerési szűrőkészlet optimalizálásával [40]. A különbség az, hogy az általunk használt módszer a hagyományos szűrőkészlet helyett spektro-temporális jellemzőket használ: úgy kezelve a spektro-temporális szűrőket, mint a neuronháló legalsó rétegét, ezzel lehetővé téve, hogy a tanítási algoritmus finomhangolja őket.

Hogy megértsük mi teszi lehetővé a módszer működését, vizsgáljuk meg a formulát amely egy neuron o kimenetét meghatározza:

$$o = a\left(\sum_{i=1}^{L} x_i \cdot w_i + b\right), \tag{5}$$

ahol x a neuron bemenete, L ezen bemenet hossza, w a súlyvektor, és b a neuronhoz tartozó ún „bias”. Az a aktivációs függvény a legtöbb esetben szigmoid, de alkalmazhatjuk akár az identitás függvényt is a feladatra. Ha ezt tesszük, és a „bias” paraméter értékét nullának választjuk, (5) egyenlőséget felírhatjuk a következő formában:

$$o = \sum_{i=1}^{L} x_i \cdot w_i. \tag{6}$$

Ha (1) egyenlőségből P ablakot és F szűrőt vektor formában írjuk fel ($\overline{P}$, $\overline{F}$ – ami csak jelölésbeli különbséget jelent), a következő alakot kapjuk:

$$o = \sum_{i=1}^{M \cdot N} \overline{F}_i \cdot \overline{P}_i. \tag{7}$$

Ha $\overline{P}$-t választjuk a neuron x bemenetének, és a $\overline{F}$-et választjuk a neuron súlyvektorának, könnyen beláthatjuk most, hogy a (6) egyenlőség a (7) egyenlőség speciális esete – és így azt is, hogy az (5) egyenlőség speciális esete az (1) egyenlőségnek. Ez azt jelenti, hogy a spektro-temporális szűrők valóban integrálhatók a neuronhálóba speciális neuronokként.

Kezdeti szűrők	Jellemzősúlyok Változatlanul	Tanítva
Random	32.96%	30.27%
2D DCT	31.19%	30.21%
Gábor Manual	32.41%	30.29%

2. táblázat. Fonémafelismerési hibbaarányok a TIMIT core teszthalmazán (20 függetlenül tanított neuronháló eredményeinek átlaga).

Először, a koncepció igazolására tiszta beszéden végeztünk kísérleteket a TIMIT adatbázis használatával. Ebben az esetben hagyományos, szigmoid aktivációs függvényt alkalmazó neuronhálókba illesztettük a jellemzőkinyerési réteget, valamint a neuronháló által felhasznált környezet hossza az időtartományban nem haladta meg a spektro-temporális ablakok hosszát. A kísérletek eredményei (lásd 2. táblázat) megmutatták, hogy az együttes optimalizálás (lásd a 2. táblázat harmadik oszlopát) valóban jobb felismerési eredményre vezet annál, mint amit akkor tudunk elérni, ha a jellemzőket rögzítjük a neuronháló tanítása előtt (lásd a 2. táblázat második oszlopát). Ezután vizsgálódásainkat kiterjesztettük nagyobb (rejtett rétegeikben több neuront tartalmazó) neuronhálókra, melyek szélesebb kontextust használtak. Ezen kísérleteket a TIMIT adatbázis mellett a Szeged magyar nyelvű híradós adatbázison is elvégeztük. Az eredmények megerősítették korábbi megállapításainkat az együttes optimalizálás hasznáról, valamint megmutatták az együttes optimalizálás előnyét zajjal szennyezett beszéd felismerésének esetére is.

Mi több, az együttes optimalizálás előnye megmutatkozott az eltérő adatbázist és nyelvet használó kísérletekben is (lásd 3. táblázat). A megadott táblázatban a hagyományos neuronhálók segítségével, az MFCC jellemzőkészlet valamint különböző – korábban részletesebben bemutatott – Gábor-szűrőkészletek (G1, G2, G3, SMK, SFFS, FFN) felhasználásával kapott eredményeket hasonlíthatjuk össze azokkal, melyeket az együttes jellemző és neuronháló optimalizálási módszerrel értünk el. A 3. táblázatból tisztán látszik, hogy a legjobb eredményeket akkor értük el, amikor az együttes optimalizálási keretrendszert alkalmaztuk, és az együtthatókat vagy a manuálisan megalkotott Gábor-szűrőkészlet alapján (Gábor Manual), vagy ugyan annak a jellemzőkészletnek a TIMIT adatbázison már korábban optimalizált változatával inicializáltuk (Gábor Manual + TIMIT).

Együttes optimalizálas	(Kezdeti) Szűrők	Jellemzősúlyok Változatlanul	Tanítva
✓	Random	26.94%	25.06%
✓	Gábor Manual	26.36%	24.75%
✓	Gábor Manual+TIMIT	25.10%	**24.64%**
	MFCC + Δs	25.03%	–
	SFFS + Δs	26.06%	–
	FFNN + Δs	26.49%	–
	G1 + Δs	25.91%	–
	G2 + Δs	27.16%	–
	G3 + Δs	36.32%	–
	SMK	26.95%	–

3. táblázat. Fonémafelismerési hibaarányok a Szeged magyar nyelvű híradós adatbázison (10 függetlenül tanított neuronháló eredményeinek átlaga).

Kezdeti Jellemzők	Eredeti keretrendszer	DRN keretrendszer	DCRN keretrendszer
Gábor	26.24%	23.37%	**22.98%**
2D DCT	26.54%	24.15%	**23.22%**
Random	26.53%	23.58%	**23.25%**

4. táblázat. Fonémafelismerési hibaarányok a TIMIT adatbázis tiszta core teszthalmazán (10 függetlenül tanított neuronháló eredményeinek átlaga).

Az együttes optimalizálás kezdeti sikere után kiterjesztettük kísérleteinket a mély tanulásra. Először – követve Glorot és Bengio munkásságát [10] – a szigmoid aktivációs függvényt alkalmazó neuronokat rectifier aktivációs függvényt alkalmazó neuronokra (ReLU) cserélve hoztunk létre mély neuronhálókat (DRN), ahol egy neuron kimenetét a következő képlet adja meg:

$$o = max\left(0, \sum_{i=1}^{L} x_i \cdot w_i + b\right). \tag{8}$$

Az így keletkezett keretrendszert ismét a TIMIT beszédadatbázison értékeltük ki. A kapott eredmények megerősítették korábbi megállapításainkat a javasolt módszerrel kapcsolatban. Mi több, az eredmények alátámasztották a DRN struktúra előnyeivel kapcsolatos feltevésünket is (lásd 4. táblázat 3. oszlopa).

Végül az együttes optimalizálásnál alkalmazott neuronhálókat az időtartományban végrehajtott konvolúció használatával egészítettük ki. Ezt oly módon tettük, hogy az – azonos súlyok használatával – feldolgozott szomszédos ablakok között minden esetben kihagytunk egy (vagy több) ablakot – megnövelve az ablakok közötti lépés nagyságát – konvolúciós hálóvá alakítva mély hálónkat. Az így kapott mély konvolúciós neuronhálót (DCRN) szitén a TIMIT adatbázison értékeltük ki. A tiszta beszédre kapott eredmények (lásd 4. táblázat) ismét azt mutatták, hogy a bevezetett csökkentette a felismerési hibaarányt. A legtöbb esetben ugyanezt tapasztaltuk zajos beszéd esetén is.

A fejezet eredményeinek tézisszerű összefoglalása

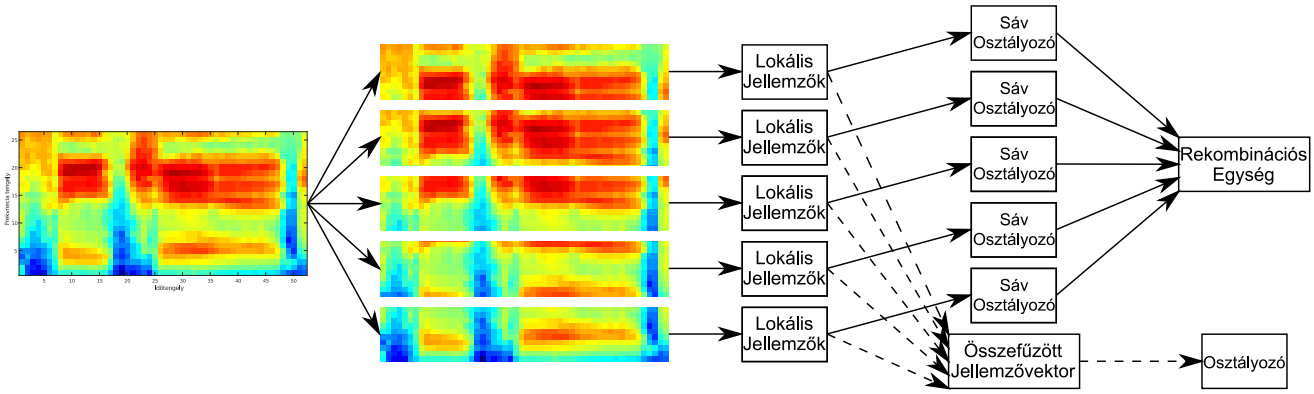
II/1. Bevezettünk egy algoritmust a spektro-temporális jellemzők és a neuronhálók együttes optimalizálására. A javasolt módszert fonémafelismerési feladatokon értékeltük ki a TIMIT angol nyelvű adatbázis, és a Szeged magyar híradós adatbázis használatával. Az eredmények megerősítették az együttes optimalizálás kivitelezhetőségét, és megmutatták, hogy jelentősen javítja a felismerési eredményeket adatbázisok közötti kísérletek esetén is (publikálva: [21, 26]).

II/2. Több, a neuronhálókkal kapcsolatos kutatásból származó eredményt építettünk be az együttes optimalizálási keretrendszerbe, úgymint a ReLUk használata, valamint a konvolúció alkalmazása. Ezen új eredmények beillesztése a keretrendszerbe kiemeli annak flexibilitását, valamint kapacitását az új eredmények befogadására. A TIMIT adatbázison végzett kísérletek eredménye megmutatta, hogy a módosításokkal a keretrendszer szignifikánsan alacsonyabb hibaarányok elérésére képes (publikálva: [22]).

4. A beszédjel többsávos feldolgozása spektro-temporális jellemzők használatával

A több forrásra épülő (multi-stream) beszédfeldolgozás folyamán a beszédjelből származó információt több, különálló forrásra bontják. Ezek a – függetlenül feldolgozott, majd később újraegyesített – források a beszédjel különböző tulajdonságait vagy aspektusait reprezentálhatják. A többsávos feldolgozás (melyet először Duchnowski írt le [7]) a több forrásra épülő feldolgozás speciális esete, ahol a különböző frekvenciatartományokat kezeljük különálló forrásokként. Ebben a megközelítésben a bemeneti jelet spektrális sávokra bontjuk, majd a sávok független feldolgozása után (mely általában részleges felismerést is tartalmaz), a belőlük származó információt egyesítve jutunk a végleges felismerési eredményekre.

A módszer alkalmazása több okra vezethető vissza, úgy mint jelfeldolgozási megfontolások, a párhuzamos számítási kapacitás kihasználásának lehetősége, az emberi beszédértéshez való hasonlóság, és a zajtűrőség (mivel szemben azzal az esettel, ahol a felismerő a teljes frekvenciatartományra támaszkodik, a többsávos esetben sávhatárolt zaj jelenlétében lesznek olyan komponensek, melyek teljesítményét a zaj nem rontja le). Ez utóbbi megfontolásokat láthattuk korábban is, a spektro-temporális feldolgozás mögötti motiváció ismertetésekor. A hasonlóság a két módszer között ezzel nem ér véget, ahogy a 4. ábrán látható sematikus reprezentációjuk is mutatja. Olyannyira hasonló a két módszer, hogy az általunk a korábban használt megközelítést (ahol a különböző frekvenciatartományokból kinyert jellemzőket egy jellemzővektorba vonjuk össze a neuronhálóval történő feldolgozás előtt) egyes munkákban a többsávos feldolgozás speciális eseteként, jellemző-rekombináció (feature recombination) módszer néven is említik [35]. Ám a 4. ábra nemcsak a két módszer hasonlóságát illusztrálja, hanem a két – hasonló elvekre (úgymint az automatikus beszédfelismerés közelítése az emberi beszédfeldolgozáshoz, valamint a zajtűrő beszéfelismerés igénye) épülő – módszer kompatibilitását is. Könnyen beláthatjuk a két módszer kompatibilitásának meglétét, amennyiben figyelembe vesszük, hogy a többsávos megoldásra úgy jutottunk, hogy a különböző frekvenciatartományokból származó jellemzőket összevonás helyett először függetlenül tanított neuronhálókba irányítottuk. Ám a nyilvánvaló kompatibilitás ellenére is, bár a két módszer hosszú idő óta használatban van, kombinációjukra kevés kísérlet történt. Azon, ritka tanulmányok, melyek a spektro-temporális jellemzőkinyerést a több forrásra épülő beszédfeldolgozással ötvözik, főleg a jellemzők tulajdonságai alapján választották szét a forrásokat [31, 47], nem pedig az alapján, hogy azok melyik frekvenciatartományból származnak.



4. ábra. A spektro-temporális feldolgozás (ahogy a második fejezetben használtuk), és a többsávos feldolgozás módszerének sematikus ábrázolása. Ahol a két módszer elválik, az előbbi folyamatát szaggatott, míg az utóbbi folyamatát folyamatos vonallal reprezentáljuk.

A többsávos feldolgozás elnevezés módszerek széles skáláját fedi le eltérő tulajdonságokkal. A felhasznált sávok száma például kettőtől akár huszonkettőig is terjedhet. Különbségek mutatkozhatnak még a sávok átfedésének kérdésében is. Míg a legtöbb esetben a felbontás úgy történik, hogy a keletkező sávok között (a mel-szűrők átfedésétől eltekintve) nincs átfedés, a felbontás átfedő ablakokkal is megtörténhet. A különböző sávok kiválasztása után további döntéseket kell meghozni sávok feldolgozásával és a kapott információ kombinációjával kapcsolatban. A sávok feldolgozása történhet GMM, vagy HMM/ANN hibrid segítségével, ám a leggyakrabban erre a célra neurális hálókat alkalmaznak [3, 32]. Továbbá döntést kell hozni arról is, hogy a keletkező információból mennyit használjunk: csupán a választott hipotézis címkéjét, esetleg a különböző hipotézisek sorrendjét, vagy a posterior valószínűségbecsléseket. Tegyük fel most, hogy a legutóbbi lehetőséget választottuk. Ez esetben is még számos lehetőségünk van a sávokból származó információ rekombinációjára, az egyszerű, előre meghatározott lineáris kombinációtól kifinomultabb módszerekig, melyek dinamikusan próbálják becsülni a különböző sávokból származó információ megbízhatóságát. Elvégezhetjük a rekombinációt továbbá újabb neurális hálókkal is [32]. Jelen munkában (ahogy a 4. ábra is sugallja) a többsávos feldolgozás egy meghatározott megközelítését használjuk. Kísérleteink során a frekvenciatartományt négytől hatig terjedő számú átfedő sávra bontjuk, és ezeket független neuronhálókkal dolgozzuk fel, majd a különböző neuronhálókból származó információt egy újabb neuronhálóban használjuk fel.

Első megközelítésben különböző neuronháló struktúrákkal kísérleteztünk a – második fejezetben bemutatott – spektro-temporális jellemzők feldolgozására, mind a külön sávok, mind pedig a rekombinációs háló esetére. Az egyik megközelítésben (MB small) a sávok feldolgozását végző háló és a rekombinációs háló is egy hagyományos MLP volt, ahol a különböző rétegekben található neuronok számát úgy határoztuk meg, hogy a tanítható paraméterek száma megegyezzen a jellemzőrekombinációs módszer (FC) esetén használtakkal. Ezen megközelítés egy módosított változatát is alkalmaztuk, ahol minden rejtett rétegben jelentősen növeltük a neuronok számát (MB big). Végül egy olyan struktúrát is létrehoztunk, ahol minden felhasznált neuronhálót mély hálóval helyettesítettünk (MB deep). Itt, a korábbi (és későbbi) fejezetekkel szemben, a mély tanulás nem ReLUk használatával, hanem Hinton és társai [13] előtanítási módszerével történt. A TIMIT adatbázison kinyert eredmények felhasználásával hasonlítottuk össze a különböző többsávos feldolgozásra épülő módszereket egymással, továbbá a többsávos feldolgozásra épülő módszereket a jellemzőrekombinációs módszerrel (lásd 6. táblázat). Úgy találtuk, hogy a többsávos megközelítés a neuronháló struktúrától függetlenül minden esetben jobb felismerési eredményre vezetett mint a jellemzőrekombinációs módszer. A legjobb eredményeket azonban a mély tanulást felhasználó többsávos feldolgozásra épülő módszerrel (MB deep) értük el. Azt is megfigyelhettük, hogy a legjobb Gábor-szűrőkkel elért eredmények minden esetben túlszárnyalták a legjobb 2D DCT jellemzőkkel elért eredményeket.

Struktúra	2D DCT	Gábor	Paraméterek száma
FC	26.85%	26.78%	~ 6
MB small	26.07%	25.45%	~ 6
MB big	24.70%	24.79%	$\sim$16
MB deep	**23.47%**	**22.81%**	$\sim$17

5. táblázat. Fonémafelismerési hibaarányok a TIMIT adatbázis tiszta core teszthalmazán (10 függetlenül tanított neuronháló eredményeinek átlaga), valamint a különböző struktúrákban használt tanítható paraméterek száma. Mind a 2D DCT, mind a Gábor-szűrők esetén a legjobb eredmény vastagon kiemelve.

Módszer	Teszthalmaz				Átlag
	A	B	C	D	
Jellemzőrekombináció	3.9%	13.4%	12.0%	28.6%	19.3%
Üvegnyak	3.7%	**12.9%**	**11.6%**	**26.2%**	**17.8%**
Ganapathy [9]	**3.0%**	**12.9%**	11.7%	27.7%	18.5%

6. táblázat. Eltérő módszerekkel elért szófelismerési hibaarányok az Aurora-4 adatbázis különböző teszthalmazain (3 függetlenül tanított neuronháló eredményeinek átlaga).

A második fázisban a – harmadik fejezetben bemutatott – együttes optimalizálási keretrendszert alkalmaztuk a frekvenciasávok feldolgozására. Az egyes sávokból kinyert információ rekombinációja ebben az esetben is rekombinációs hálóval történt. Ezúttal azonban ennek bemenetét az egyes frekvenciasávokat felhasználó hálókba (a kimeneti réteg elé) helyezett ún. üvegnyak (bottleneck) réteg kimenete szolgáltatta. Az így kapott többsávos üvegnyak módszert az Aurora-4 angol nyelvű beszédadatbázison értékeltük ki, az ARMA jellemzők [9] felhasználásával. Ebben az esetben is azt a feladatot kíséreltük meg megoldani, ahol a tanítás során csupán tiszta beszédet használhatunk. A kapott eredményeket (lásd 6. tábázat) négy csoportba osztottuk. Az A teszthalmaz alatt azok a szófelismerési hibaarányok szerepelnek, melyeket a tanító halmaz rögzítésénél is alkalmazott Sennheiser mikrofonnal felvett tiszta beszéd esetére kaptunk. A B teszthalmazra ugyan ezzel a mikrofonnal felvett, de különböző zajokkal szennyezett beszéd esetére kapott szófelismerési hibaarányokat jelentettünk. A C teszthalmaz az A teszthalmazhoz hasonlóan tiszta beszédet tartalmaz, de itt a beszéd rögzítése más mikrofonokkal történt. Végül a D teszthalmazba tartozó mondatok szintén ezekkel a mikrofonokkal kerültek rögzítésre, ám az ide tartozó mondatok (a B teszthalmazhoz hasonlóan) különböző zajok hozzáadásával készültek. Az összehasonlítás kedvéért a 6. táblázat legalsó sora tartalmazza a legjobb eredményt, amit az adott feladatra találtunk [9]. Az eredmények azt mutatják, hogy a többsávos feldolgozás az együttes optimalizálással kombinálva is javítja annak eredményeit. Továbbá azt tapasztaltuk, hogy a többsávos feldolgozás az ARMA spektrogrammal kombinálva versenyképes szófelismerési hibaarányokat eredményez.

A fejezet eredményeinek tézisszerű összefoglalása

III/1. A többsávos feldolgozás és a spektro-temporális jellemzőkinyerés kompatibilitását felhasználva bemuttatunk egy módszert a két megközelítés kombinálására. Tettük ezt a második fejezetben bemutatott spektro-temporális jellemzők felhasználásával. Megmutattuk a TIMIT adatbázison, hogy a többsávos feldolgozás jobb eredményeket ad tiszta, valamint zajjal szennyezett beszéd esetén is (publikálva: [25]).

III/2. Szintén bemutattunk egy módszert, ami a többsávos feldolgozást az együttes optimalizációs keretrendszerrel ötvözi. A javasolt módszert az Aurora-4 adatbázison értékeltük ki, kizárólag tiszta tanító adatok felhasználásával. Megmutattuk, hogy az így kapott módszer jobban teljesít, mint a korábban ismertetett együttes optimalizálási módszer önmagában. Mi több, a többsávos feldolgozással elért eredmények az eddigi legjobb eredmények között vannak, melyeket az adott feladatra publikáltak (publikálva: [23]).

Módszer	PER
Baby és tsai. [2]	19.6%
Plahl és tsai. [37]	19.1%
Tóth [44]	18.7%
Jelen tanulmány	*18.5%*
Graves és tsai. [11]	17.7%
Tóth [45]	**16.7%**

7. táblázat. Fonémafelismerési hibaarányok (PER) az irodalomban a TIMIT core teszthalmazra. A legjobb eredmény vastagon kiemelve. Az általunk elért eredmény dőlt betűvel szedve.

5. Sávkiejtés (band dropout)

Ezen tanulmány nagy részében az volt a legfontosabb kérdés egy új módszer ismertetésénél, hogy az milyen hatással van az elért eredményeinkre: alacsonyabb hibaszázalékokat kapunk-e az általunk korábban elért eredményeknél, vagy sem. Kutatásainkat azonban nem vákumban végezzük, hanem egy olyan környezetben, ahol egyre újabb eredményeket tesznek közzé minden adatbázisra. Így azt is fontos megvizsgálni, hogy az ismertetett módszerek teljesítménye hogy viszonyul az irodalomban található hasonló módszerek eredményeihez.

A tanulmány ezen részében a korábban ismertetett módszereket ötvöztük és fejlesztettük tovább, hogy megmutassuk, ezek a módszerek képesek versenyképes eredményeket produkálni. Mielőtt azonban új módszert vezettünk volna be, először megvizsgáltuk az eredeti együttes optimalizálási keretrendszert, hogy lássuk hogyan tudnánk tovább javítani rajta, és hogy lássuk képes-e önmagában olyan eredményeket produkálni, melyek összevethetők az irodalomban találtakkal. Ehhez először finomhangoltuk az együttes optimalizálás korábban általunk nem vizsgált paramétereit a TIMIT adatbázison végzett kísérletek alapján. Azután, a Δ és $\Delta\Delta$ együtthatók korábbi (lásd második fejezet) sikeres alkalmazása által motiváltan bevezettük a deltához hasonló együtthatók használatát az együttes optimalizálási keretrendszerbe. A javasolt módosítások hatását a TIMIT, valamint az Aurora-4 angol beszédadatbázisokon értékeltük ki. Az eredmények azt mutatták, hogy mindkét módosítás szignifikánsan javította a felismerési eredményeket. A legjobb elért eredményeket aztán összehasonlítottuk az irodalomban talált eredményekkel is, a TIMIT (lásd 7. táblázat), majd az Aurora-4 adatbázison (lásd 8. táblázat). Azt találtuk, hogy bár az általunk elért eredményeknél vannak jobbak, azok versenyképesek a hasonló módszerek eredményeivel.

Módszer	WER
Chang és Morgan [5]	16.6%
Seltzer és tsai. [43]	12.4%
Martinez és tsai. [30]	12.3%
Baby és tsai. [1]	11.9%
Jelen tanulmány	*11.6%*
Narayanan és Wang [34]	11.1%
Rennie és tsai. [39]	**10.3%**

8. táblázat. Szófelismerési hibaarányok (WER) az irodalomban az Aurora-4 adatbázison, vegyes tanítási adatok esetén. A legjobb eredmény vastagon kiemelve. Az általunk elért eredmény dőlt betűvel szedve.

Módszer	WER
CNN és ARMA jellemzők, sávkiejtés (band dropout)	**16.0%**
Többsávos CNN és ARMA jellemzők	17.8%
DNN és ARMA jellemzők, valamint DCT [9]	18.5%
DNN és DNN speech enhancement of FBANK [15]	17.5%
DNN és Spectral masking [29]	22.8%
CNN és PNS jellemzők valamint Gabor Filter Kernels [5]	22.9%
DNN és Exemplar Based Enhancement [1]	26.8%

9. táblázat. Az ARMA jellemzőkkel használt sávkiejtési módszer eredményeinek összehasonlítása hasonló módszerekkel az Aurora-4 adatbázison, tiszta tanítási adatok felhasználásának esetére.

Végül keretrendszerünket kiegészítettük a bemeneti kiejtés (input dropout [14]) által inspirtál sávkiejtés (band dropout) módszerrel. Ebben a módszerben a bemeneti jellemzők független kiejtése helyett azonban a kiejtés olyan módszerét javasoljuk, amelyet teljes frekvenciasávokon alkalmazunk. Feltételezésünk szerint ezáltal rászoríthatjuk a hálót, hogy ne támaszkodjon a teljes frekvenciatartományra, így érve el annak jobb zajtűrését. A kiértékelést az Aurora-4 adatbázison végeztük el tiszta, valamint zajos tanítás esetén. A legversenyképesebb eredményt összehasonlítottuk az irodalomban talált újabb eredményekkel (lásd 8. táblázat), ahol azt találtuk, hogy módszerünk jobban teljesített, mint az általunk az irodalomban talált hasonló módszerek.

A fejezet eredményeinek tézisszerű összefoglalása

IV/1. Bemutattuk két módosítását az együttes neuronháló optimalizálási és jellemző kinyerési keretrendszernek, és kiértékeltük őket a TIMIT valamint Aurora-4 angol nyelvű beszédadatbázisokon. Azt találtuk, hogy mind a két módosítás szignifikánsan javította a felismerési eredményeket, és általuk olyan eredményeket kaptunk, melyek versenyképesek az irodalomban talált hasonló módszerek eredményeivel (publikálva: [24]).

IV/2. Bemutattunk egy új bemenetkiejtési (dropout) módszert, amely kifejezetten hasznosnak bizonyult a CNN alapú akusztikus modellezéssel ötvözve. A módszer hatékonyságát az Aurora-4 adatbázison demonstráltuk különböző tanítási forgatókönyvek és különböző spektrális reprezentációk esetére (publikálva: [27]).

	[19]	[20]	[21]	[22]	[23]	[24]	[25]	[26]	[27]
I/1.	•	•							
I/2.		•							
I/3.								•	
II/1.			•					•	
II/2.				•					
III/1.							•		
III/2.					•				
IV/1.						•			
IV/2.									•

10. táblázat. A kapcsolat a tézispontok és a szerző publikációi között.

Hivatkozások

[1] Baby, D., Gemmeke, J. F., Virtanen, T., and Van hamme, H. Exemplar-based speech enhancement for Deep Neural Network based automatic speech recognition. In *Proc. ICASSP* (2015), pp. 4485–4489.

[2] Baby, D., and Van hamme, H. Investigating modulation spectrogram features for Deep Neural Network-based automatic speech recognition. In *Proc. Interspeech* (2015), pp. 2479–2483.

[3] Bourlard, H., and Dupont, S. Subband-based speech recognition. In *Proc. ICASSP* (1997), pp. 1251–1254.

[4] Bouvrie, J., Ezzat, T., and Poggio, T. Localized spectro-temporal cepstral analysis of speech. In *Proc. ICASSP* (2008).

[5] Chang, S.-Y., and Morgan, N. Robust CNN-based speech recognition with Gabor filter kernels. In *Proc. Interspeech* (2014), pp. 905–909.

[6] Chi, T., Ru, P., and Shamma, S. A. Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am. 118*, 2 (2005), 887–906.

[7] Duchnowski, P. *A New Structure for Automatic Speech Recognition.* PhD thesis, MIT, 1993.

[8] Ezzat, T., Bouvrie, J. V., and Poggio, T. A. Spectro-temporal analysis of speech using 2D Gabor filters. In *Proc. Interspeech* (2007), pp. 506–509.

[9] Ganapathy, S. Robust speech processing using ARMA spectrogram models. In *Proc. ICASSP* (2015), pp. 5029–5033.

[10] Glorot, X., and Bengio, Y. Understanding the difficulty of training Deep Feedforward Neural Networks. In *Proc. AISTATS* (2010).

[11] Graves, A., Mohamed, A., and Hinton, G. E. Speech recognition with Deep Recurrent Neural Networks. In *Proc. ICASSP* (2013), pp. 6645–6649.

[12] Heřmanský, H., and Sharma, S. TRAPS-classifiers of temporal patterns. In *Proc. ICSLP* (1998), pp. 1003–1006.

[13] Hinton, G., Deng, L., Yu, D., Abdel-rahman, M., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Dahl, G., and Kingsbury, B. Deep Neural Networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine 29*, 6 (2012), 82–97.

[14] Hinton, G., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Improving Neural Networks by preventing co-adaptation of feature detectors. *CoRR abs/1207.0580* (2012).

[15] Jun, D., Qing, W., Tian, G., Yong, X., Li-Rong, D., and Chin-Hui, L. Robust speech recognition with speech enhanced Deep Neural Networks. In *Proc. Interspeech* (2014), pp. 616–620.

[16] Kanedera, N., Arai, T., Heřmanský, H., and Pavel, M. On the relative importance of various components of the modulation spectrum for automatic speech recognition. *Speech Communication 28*, 1 (1999), 43–55.

[17] Kleinschmidt, M. *Robust Speech Recognition Based on Spectro-Temporal Processing*. PhD thesis, Carl-von-Ossietzky Universitt Oldenburg, 2002.

[18] Kleinschmidt, M., and Gelbart, D. Improving word accuracy with Gabor feature extraction. In *Proc. ICSLP* (2002).

[19] Kovács, G., and Tóth, L. Localized spectro-temporal features for noise-robust speech recognition. In *Proc. ICCCONTI* (2010), pp. 481–485.

[20] Kovács, G., and Tóth, L. Phone recognition experiments with 2D-DCT spectro-temporal features. In *Proc. SACI* (2011), pp. 143–146.

[21] Kovács, G., and Tóth, L. The joint optimization of spectro-temporal features and Neural Net classifiers. In *Proc. TSD* (2013), pp. 552–559.

[22] Kovács, G., and Tóth, L. Joint optimization of spectro-temporal features and Deep Neural Nets for robust automatic speech recognition. *Acta Cybernetica 22*, 1 (2015), 117–134.

[23] Kovács, G., and Tóth, L. Multi-band noise robust speech recognition using Deep Neural Networks (in Hungarian). In *Proc. MSZNY* (2016), pp. 287–294.

[24] Kovács, G., and Tóth, L. Optimisation of a spectro-temporal feature selection method integrated in Deep Neural Networks (in Hungarian). In *Proc. MSZNY* (2017), pp. 158–169.

[25] Kovács, G., Tóth, L., and Grósz, T. Robust multi-band ASR using Deep Neural Nets and spectro-temporal features. In *Proc. SPECOM* (2015), pp. 386–393.

[26] Kovács, G., Tóth, L., and Van Compernolle, D. Selection and enhancement of Gabor filters for automatic speech recognition. *IJST 18*, 1 (2015), 1–16.

[27] Kovács, G., Tóth, L., Van Compernolle, D., and Ganapathy, S. Increasing the robustness of CNN acoustic models using autoregressive moving average spectrogram features and channel dropout. *Pattern Recognit. Lett.* (2017).

[28] Lamel, L. F., Kassel, R., and Seneff, S. Speech database development: design and analysis of the acoustic-phonetic corpus. In *Proc. DARPA Speech Recognition Workshop, Report no. SAIC-86/1546* (1986).

[29] Li, B., and Sim, K. C. Improving robustness of Deep Neural Networks via spectral masking for automatic speech recognition. In *Proc. ASRU* (2013), pp. 279–284.

[30] Martínez, A. M. C., Moritz, N., and Meyer, B. T. Should Deep Neural Nets have ears? The role of auditory features in deep learning approaches. In *Proc. Interspeech* (2014), pp. 2435–2439.

[31] Mesgarani, N., Thomas, S., and Heřmansky, H. A multistream multiresolution framework for phoneme recognition. In *Proc. Interspeech* (2010), pp. 318–321.

[32] Mirghafori, N. *A Multi-Band Approach to Automatic Speech Recognition.* PhD thesis, International Computer Science Institute, 1999.

[33] Morgan, N., and Bourlard, H. Continuous speech recognition using multilayer perceptrons with Hidden Markov Models, 1990.

[34] Narayanan, A., and Wang, D. Joint noise adaptive training for robust automatic speech recognition. In *Proc. ICASSP* (May 2014), pp. 2504–2508.

[35] Okawa, S., Bocchieri, E., and Potamianos, A. Multi-band speech recognition in noisy environments. In *Proc. ICASSP* (1998), pp. 641–644.

[36] Parihar, N., and Picone, J. DSR front end LVCSR evaluation. Aurora Working Group AU/384/02, Institue for Signal and Information Processing, December 2002.

[37] Plahl, C., Sainath, T. N., Ramabhadran, B., and Nahamoo, D. Improved pre-training of deep belief networks using sparse encoding symmetric machines. In *Proc. ICASSP* (2012), pp. 4165–4168.

[38] Pudil, P., Novovičová, J., and Kittler, J. Floating search methods in feature selection. *Pattern Recogn. Lett. 15*, 11 (Nov. 1994), 1119–1125.

[39] Rennie, S. J., Dognin, P. L., Cui, X., and Goel, V. Annealed dropout trained maxout networks for improved LVCSR. In *Proc. ICASSP* (2015), pp. 5181–5185.

[40] Sainath, T. N., Kingsbury, B., rahman Mohamed, A., and Ramabhadran, B. Learning filter banks within a Deep Neural Network framework. In *Proc. ASRU* (2013), pp. 297–302.

[41] Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L., Roomi, B., and Hall, P. English conversational telephone speech recognition by humans and machines. *CoRR abs/1703.02136* (2017).

[42] Schädler, M. R., Meyer, B. T., and Kollmeier, B. Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition. *J. Acoust. Soc. Am. 131* (2012), 4134–4151.

[43] Seltzer, M. L., Yu, D., and Wang, Y. An investigation of Deep Neural Networks for noise robust speech recognition. In *Proc. ICASSP* (May 2013), pp. 7398–7402.

[44] Tóth, L. Convolutional Deep Rectifier Neural Nets for phone recognition. In *Proc. Interspeech* (2013), pp. 1722–1726.

[45] Tóth, L. Combining time- and frequency-domain convolution in Convolutional Neural Network-based phone recognition. In *Proc. ICASSP* (2014), pp. 190–194.

[46] Tóth, L., and Grósz, T. A comparison of Deep Neural Network training methods for large vocabulary speech recognition. In *Proc. TSD* (2013), pp. 36–43.

[47] Zhao, S. Y., Ravuri, S. V., and Morgan, N. Multi-stream to many-stream: using spectro-temporal features for ASR. In *Proc. Interspeech* (2009), pp. 2951–2954.