Eredmények a magyar nyelvű beszédfelismerési konfidencia-becslésben

Tarján Balázs, Györki Milán, Mihajlik Péter és Gordos Géza

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék, Távközlési és Beszéd-jelfeldolgozási Laboratórium
{btarjan, gyorki, mihajlik}@tmit.bme.hu

Kivonat: A beszédfelismerési konfidencia-becslés célja, minden felismeréshez egy megbízhatósági mérőszámot rendelni. Valódi konfidenciáról viszont csak akkor beszélhetünk, ha a mérőszám jól közelíti a felismerési valószínűséget. Megbízható becslését számos gyakorlati feladat igényli, hiszen nem működhet egy felismerő optimálisan, ha nem képes különbséget tenni biztos és bizonytalan felismerés között. Kísérleteink során sikerült olyan számítási módszert kidolgoznunk, mellyel a konfidencia jól közelíthetővé vált. Cikkünkben összefoglaljuk módszerünk elméleti alapjait, a gyakorlati rendszer felépítését és a rajta elvégzett méréseink eredményeit.

1 Bevezetés

A beszéd gépi felismerésénél sohasem lehetünk biztosak abban, hogy adott felismerési eredmény nem csak optimális, de helyes is egyben. Ezért a gyakorlati alkalmazásoknál jelentős segítséget nyújthat egy olyan eljárás, amely minden felismeréshez egy megfelelő megbízhatósági mérőszámot, azaz **konfidenciát** rendel. Konfidencia birtokában lehetőségünk nyílik egyetlen valószínűségi mérőszámban megragadni mindazt a bonyolult folyamatot, mely az emberi bemondáshoz, egy lehetséges felismert szót rendel.

Megbízható mérőszám előállításának feltétele az optimális felismerő hálózat és a hozzá illeszkedő pontos becslés. Tehát két területen van lehetőség az előrelépésre. Egyfelől különböző felismerési konfigurációk alkalmazásával, másfelől a konfidencia-számítás módszerének változtatásával javíthatunk a becslésen. A cikkünk alapjául szolgáló kísérletsorozatban ezek számos változatát vizsgáltuk pontosságuk alapján, és minden esetben a legjobban teljesítőket emeltük ki. Az így adódott összesen négy kísérleti elrendezés szolgáltatja mérési eredményeinket.

Cikkünk elején összefoglaljuk a konfidencia közelítésének általános alapelveit, és áttekintjük az általunk használt kísérleti elrendezést és annak fontos elemeit. Kitérünk a kísérleti adatbázisra, majd cikkünk második felében a konkrét implementációkon végzett vizsgálatok eredményeit és az azokból levonható következtetéseket ismertetjük.

2 Az alkalmazott konfidencia-becslési alapelvek

2.1 Mintaillesztés a normál nyelvtani hálózathoz

A beszédfelismerés első szakaszában történik a lényegkiemelés. Itt az emberi beszéd időfüggvényének minden 10msec hosszú szakaszához rendelünk egy leíró vektort. Ezt nevezzük jellemzővektornak.

Második lépésben történik a mintaillesztés **[6]**, ennek során a jellemzővektorok sorozatához (jelölése: *O*) hozzárendeljük a szótárban található felismerhető eredmények (jelölése: *W*) közül a legjobban illeszkedőt. Ez a szótár két előzetesen betáplált modellje alapján működik. Az egyik az úgynevezett akusztikai modell, mely az adott nyelvre jellemző fonémák jellemzővektor eloszlását rögzíti. Segítségével közelítőleg megkapjuk, hogy egy szótári elemhez milyen valószínűséggel rendelődik egy jellemzővektor sorozat ($P(O \mid W)$).

A másik fontos elem a nyelvi modell. Ez a felismerhető szavak egymás közti viszonyát leíró egyszerű vagy igen bonyolult hálózat, amiben az élek szavakat kötnek össze, és minden átmenethez rendelődik egy átmeneti valószínűségi érték. A szavak helyébe helyébe fonémák és az azokhoz tartozó akusztikai modellből származó eloszlások helyettesíthetőek be. Ilyen módon az akusztikai modell szervesen beágyazódik a nyelvi modellbe. Ez a modell folyamatos felismerésnél általában szöveges tanítás útján jön létre, ilyenkor valószínűségével szerepelnek benne ($P(W)$). Parancsszóvezérlésnél pedig egy lista megadásával rendelhetünk 1-es értéket az elfogadott szavakhoz.

Összefoglalóan ezt a hálózatot nevezzük a normál **nyelvtan**hoz (grammar) tartozó felismerési hálózatnak. Amikor a jellemzővektorhoz legjobban illeszkedő nyelvtanban rögzített elemet keressük ($\hat{W}$), akkor tulajdonképpen $P(W \mid O)$ valószínűséget kívánjuk W argumentuma mentén maximalizálni (**1**). **[2]**

$$\hat{W} = \arg\max_{W} P(W \mid O) \tag{2}$$

Ezzel ekvivalens problémához jutunk, ha alkalmazzuk a Bayes – formulát (**2**).

$$\hat{W} = \arg\max_{W} \frac{P(W) * P(O \mid W)}{P(O)} \tag{2}$$

Az utóbbi, már átalakított képlet tovább egyszerűsíthető figyelembe véve, hogy P(O) értéke független W-től (**3**).

$$\hat{W} = \arg\max_{W} P(W \mid O) = \arg\max_{W} P(W) * P(O \mid W) \tag{3}$$

Tehát a jellemzővektor sorozatot végigfutatva a nyelvtani hálózaton lépésről lépésre vehetők az illeszkedési, illetve átmeneti valószínűségek szorzatai. A legvalószínűbb felismerést az az útvonalvonal jelöli ki, mely mentén ez a szorzat maximális. Ezt a szorzatot nevezzük a felismeréshez rendelt **hasonlósági mérték**nek.

2.2 A konfidencia-becslése

A bemondás normál nyelvtani hálózathoz illesztésével megkereshetjük a legvalószínűbb illeszkedő szótári eleme(ke)t, ám az így kapott hasonlósági mértékből nem lehet következtetni az illeszkedés valószínűségére. Ennek triviális oka, hogy a hasonlósági mérték függ a végigjárt állapotok számától. Emellett, mivel minden bemondáshoz csak egyetlen legjobban illeszkedő szó rendelődik, arról sem rendelkezünk információval, hogy vajon a többi lehetséges felismerés közül lényegesen kiemelkedik-e a legjobbnak választott.

Tehát valódi konfidencia számításához vissza kell nyúlni eredeti valószínűségi képletünkhöz **[1]** illetve annak Bayes - formulával átalakítottjához (**4**).

$$P(W \mid O) = \frac{P(W) * P(O \mid W)}{P(O)} \tag{4}$$

Az egyenlet bal oldala pontosan a keresett felismerési valószínűséget fejezi ki. A jobb oldalon lévő tagok közül *P(W)* illetve *P(O|W)* számíthatóságát már korábban láttuk, így *P(O)* megfelelő közelítése jelentheti a megoldást. Az ilyen *P(O)* jellemzővektor sorozat előfordulási valószínűségeket azonban megkaphatjuk *P(O|W)* együttes eloszlásának peremeloszlásaként, mint azt az alábbi képlet is kifejezi (**5**).

$$P(O) = \sum_{W} P(W) * P(O \mid W) \tag{5}$$

A beszédfelismerésben alkalmazott nyelvtani hálózatokkal meg tudjuk határozni ennek az összegnek egy tagját, ezt használtuk ki a mintáink illesztésénél, ám a véges szótár nem alkalmas ezen végtelen összeg közelítésére. Szóba jöhetne az úgynevezett N-best módszer **[3]** alkalmazása, amikor a nyelvtani hálózat N legjobb (N>1) illeszkedéséhez tartozó hasonlósági mértékek összegével közelítjük a sort. Izolált szavas felismerésnél azonban ez nem alkalmazható, mert szótáron kívüli bemondás esetén előálló alacsony hasonlósági mértékek közül az N legnagyobb sem dominál, így azok összege és a valódi sorösszeg jelentősen eltér.

Ezért felmerül az igény egy felismerési hálózat összeállítására, mely minden jellemzővektor sorozathoz jól illeszkedik. Egy ilyen **mindenhez illeszkedő hálózat**ban a legjobb illeszkedéséhez tartozó hasonlósági mérték bemondástól függetlenül mindig kiemelkedő. Ezáltal domináns tagja is a P(O)-t meghatározó sornak, azaz vele a sorösszeg is jól közelíthető (**6**).

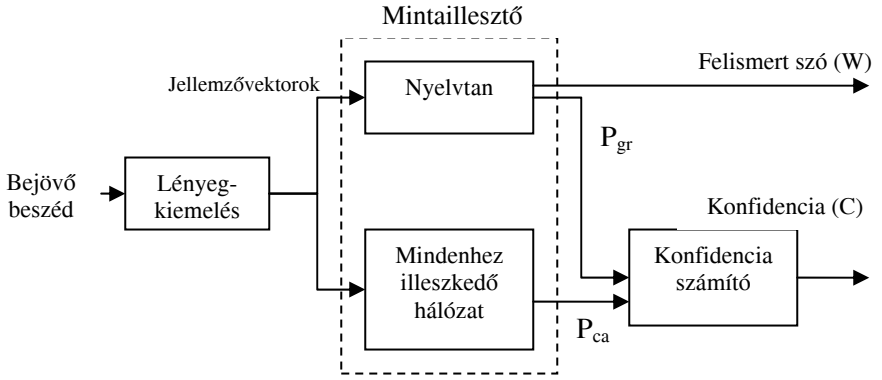
$$P(O) \cong P(W') * P(O \mid W') \tag{6}$$

A mindenhez illeszkedő hálózat, tehát becsülhetővé teszi a P(O) értékét, ami a konfidencia képletének (4) utolsó, idáig ismeretlen tagja volt.

3 A kísérleti elrendezés

Az illeszkedés mértékét a hasonlósági mérték fejezi ki, melyet a normál nyelvtannal végzett mintaillesztés szolgáltat, ám nagysága természetszerűen függ a bemondás jellemzővektor-számától, így magában nem használható konfidencia számítására.

Esetünkben a normál nyelvtani hálózat (nyelvtan) mellett egy különleges, mindenhez illeszkedő felismerési hálózatot is használunk. Konfidencia számításához e két felismerési hálózatot kötjük párhuzamosan, így egy bemondás kiértékelése mindkettőn megtörténik. (**1. ábra**)



1. ábra A párhuzamosan alkalmazott két felismerési hálózat blokkvázlata

A módszer lényege, hogy a nyelvtanhoz való illeszkedést kifejező hasonlósági mértéket (P_{gr}) hasonlítjuk össze a mindenhez illeszkedő hálózaton mért hasonlósági mértékkel (P_{ca}). E két mennyiség aránya alapján a konfidencia már becsülhető (**7**). P_{gr} és P_{ca} hányadosának értéke azonban rendszerint nem esik a kívánatos [0 - 1] valószínűségi tartományba, így „a", „b" értékekkel transzformáljuk, majd az esetlegesen mégis kívül eső elemeket 0 alatt mindig 0-nak, 1 fölött mindig 1-nek tekintjük.

$$C = \begin{cases} C_0 = f\left(\frac{P_{gr}}{P_{ca}}\right) = b * \left(\lg P_{gr} - \lg P_{ca} + a\right), & 0 \le C_0 \le 1 \\ 0, & C_0 < 0 \\ 1, & C_0 > 1 \end{cases} \quad (7)$$

Vizsgálataink során kétféle típusú mindenhez illeszkedő hálózattal dolgoztunk, egy úgynevezett **fonéma-bigram** modellre épülővel valamint egy **multi-Gauss** típusúval. Emellett mindkét esetben kétféle konfidencia variánst használtunk, a pontosabb becslés érdekében. Az első, alap variáns az imént ismertetett módón számítódik, a második, normált variáns is csak annyiban tér el ettől, hogy a logaritmikus valószínűségek különbségét osztjuk a bemondás jellemzővektor számával, és ehhez az értékhez illesztjük „a"-t illetve „b"-t.

4 A mindenhez illeszkedő felismerési hálózatok

Mint láttuk, P(O) közelítéséhez szükségünk van egy olyan felismerő hálózatra, mely minden magyar nyelvű bemondáshoz jól illeszkedik. Így, ha a normál nyelvtanhoz

illeszkedő a bemondás, akkor közel ugyanolyan hasonlósági mértéket rendel hozzá, mint a nyelvtani hálózat. Ha azonban a normál felismerési hálózathoz nem illeszkedik a bemondás, akkor ott ugyan alacsonyabb lesz a hasonlósági mérték, de minden elfogadó hálózatnál a jó illeszkedés miatt marad a magas hasonlósági mérték. A kérdés tehát, hogyan készíthetünk ilyen hálózatokat.

4.1 Multi-Gauss hálózat

Mivel a beszédhangmodellek tanítása során a hozzájuk tartozó rejtett Markov-modell állapotokat eleve GMM (Gauss Mixure Model)-lel jellemezzük, kézenfekvő, hogy az összes beszédhangmodell Gauss-komponenseit összefésüljük, és így egy olyan „meta beszédhangot" használjunk, mely mindenbeszédhangra illeszkedik. Ugyanezt tettük a szünetmodellekkel, és a két „meta modellt" párhuzamosan kapcsolva és visszahurkolva elkészült az általunk multi-Gauss-nak nevezett mindent felismerő hálózat.

A tapaszatalataink szerint az összes Gauss függvény használata messzemenően redundáns modell, ezért a [7]-szerinti „greedy"-algoritmussal a Gauss függvények számát a kezdeti közel 3000-ről 60-ra csökkentettük.

4.1 Fonéma-bigram hálózat

Az alternatív – jóval nagyobb számításigényű – mident felismerő hálózatunk a fontikailag változatos szöveganyagon tanított fonéma-bigram model volt, amiben szavakon átívelő „cross-word" trifón beszédhangmodelleket használtunk.

5 A vizsgálati adatbázis és az alkalmazott beállítások

5.1 A beszédadatbázisok

Tanítás és tesztelés céljára a legnagyobb magyar telefonos adatbázisokat használtunk (MTBA, a Besztel, a SpeechDat és a Tesztel) **[5]**. Ezek az adatbázisok elsősorban olvasott beszédet, valamint kisebb arányban spontán bemondásokat is tartalmaznak. Az első három adatbázis lényegében ugyanarra a szövegkorpuszra épül, és mindegyiknek az általunk elérhető része 500 beszélőtől tartalmaz hanganyagot. A Tesztel adatbázis 100 beszélős, és jellegzetessége, hogy szándékosan nagy és természetes háttérzajban felvett bemondásokat tartalmaz. Az adatbázisokban a vonalas és mobil telefonos felvételek összességében körülbelül ugyanolyan számban képviseltetik magukat.

5.2 Tanító- és teszthalmazok

Tanítás céljára az MTBA adatbázis 500 beszélőjének azon fonetikailag változatos mondatait és szavait jelöltük ki, melyek nem „o", és „z" jelzésűek, azaz nem tartal-

maznak tulajdonneveket és bizonyos típusú mondatokat. Ez összességében 6000 tanító felvételt eredményezett.

A 2475 felvételes izolált szavas teszthalmazunkat úgy állítottuk össze, hogy ne legyen átfedés a tanítóhalmazban szereplő beszélőkkel. Így a tanításnál fel nem használt felvételek összesen 170 beszélőtől (Besztel 100, Speechdat 50, Tesztel 20) kerültek a teszthalmazba.

5.3 Beszédfelismerési paraméterek, beállítások

Lényegkiemelési paraméterekként a bemenő beszédjelből MFCC (Mel Frequency Cepstral Coefficients) 38 dimenziós vektorokat képeztünk statikus energiát nem használva, de egyébként dinamikus Delta és Delta-Delta értékeket is számítva. Mind tanítás, mind tesztelés során alkalmaztuk a vak csatornakiegyenlítés módszerét.

Akusztikus modellként balról-jobbra struktúrájú, háromállapotú rejtett Markov-modelleket használtunk mind a monofón, mind a trifón beszédhangmodellek esetén. Állapotonként maximum 10 Gauss függvényből álló folyamatos megfigyelési sűrűségfüggvényeket használtunk. Szóhatárokon átívelő, azaz „cross-word" trifón modelleket alkalmaztunk.

6 Vizsgálat ROC görbék segítségével

6.1 A használt mérőszámok

Kísérleteink kiindulási alapja az egyszerű konfidencia-becslő ($\log P_{gr}$ és $\log P_{ca}$ különbsége). Ez az érték bár nem tartalmazza a helyes felismerés valószínűségét már alkalmas arra, hogy segítségével döntsünk a felismert szavak elfogadásáról, azaz beengedési küszöbként használható. Ezen küszöb változtatásával munkapontok jelölhetőek ki. A munkapont kijelölése után a küszöb alá került (elutasított), de egyébként helyesen felismert szavak számának (*KaH*), és összes helyesen felismert szó számának (*H*), valamint a küszöb fölötti félreismert (*KfF*) és összes félreismert szó számának (*F*) ismeretében a munkapontot jellemző két mennyiség definiálható (**8**).

$$FRR = \frac{KaH}{H} * 100[\%] \quad FAR = \frac{KfF}{F} * 100[\%] \tag{8}$$

A tévesen elutasított szavak arányát fejezi ki a False Rejection Rate (FRR), illetve a tévesen beengedettekét a False Acceptance Rate (FAR). A felismerés minősége általánosan jellemezhető az úgynevezett **Equal Error Rate** (**EER**) **[4]** mérőszámmal, ami FAR és FRR értéke annál a küszöbnél, ahol a két mennyiség megegyezik.

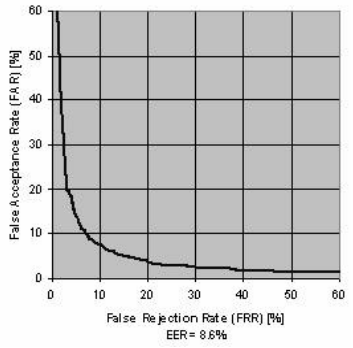
Az EER csökkentése egy fontos, de nem mindenek előtt való feladat a felismerési kísérleteknél. Ennek oka, hogy eltérőek lehetnek a felismerővel szemben támasztott elvárások. Például egy telefonközpontban használt dialógus rendszert érdemes alacsony FRR -tel rendelkező munkaponton üzemeltetni (pl.: 5%), mert nem kívánunk a helyes bemondásokat felesleges elutasításával kellemetlenséget okozni. Ilyen esetek-

ben fontosabb paraméter a kötött FRR -hez tartozó munkaponti FAR, ami nem feltétlenül a legkisebb EER -tel rendelkező felismerőben a legkedvezőbb.

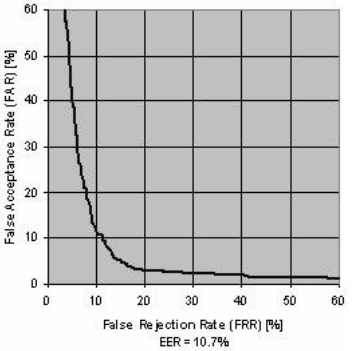
6.2 Az ROC görbék kiértékelése

Vizsgálatainkban két kísérleti elrendezést alkalmaztunk, ezeket az alkalmazott mindenhez illeszkedő hálózat különböztette meg (fonéma-bigram, multi-Gauss). Mindkét elrendezésre a korábban ismertetett két konfidencia-számító módszert alkalmaztuk, azonban számításuknál különbözőképpen definiáljuk az egyszerű konfidencia-becslőt. Alap esetben ez a hasonlósági mértékek logaritmusának különbsége (EKB_{alap}), normált esetben ugyanez a jellemzővektorok számával normálva ($EKB_{normált}$).

Az egyes munkapontok mentén összetartozó FRR és FAR értékeket egy grafikonon ábrázolva jutunk a **Receiver Operating Characteristic** (**ROC**) **[2]** görbéhez. Az ROC görbe jól szemlélteti a felismerő azon képességét, hogy milyen mértékben képes szótárban nem rögzített szavakat a szótári szavaktól szeparálni Vizsgáljuk először a két alapesetet! (**2. ábra**)



2.1. ábra Fonéma-bigram mindenhez illeszkedő hálózat, alap becslés



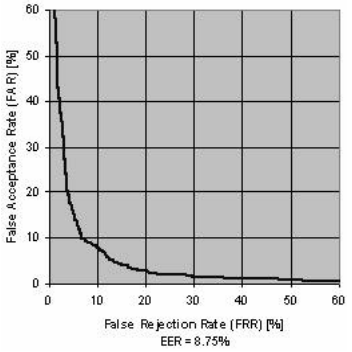
2.2. ábra Multi-Gauss mindenhez illeszkedő hálózat, alap becslés

A mindenhez illeszkedő hálózatok cseréje látványos változást idéz elő az ROC görbék alalkulásában. Multi-Gauss hálózat használtánál (**2.2. ábra, 10.7%**) lényegesen nagyobb EER mérhető, mint fonéma-bigram (**2.1. ábra, 8.6%**) modell esetén, de ennél is szembeötlőbb a görbe viselkedésének különbsége FAR tengely közelében. Ez azért különösen lényeges része az ábrának, mert a már korábban említett és eltertjedten használt FRR érzékeny alkalmazásokat az alacsony FRR értékkel rendelkező, tehát FAR tengely közelében lévő munkapontokon érdemes működteni.

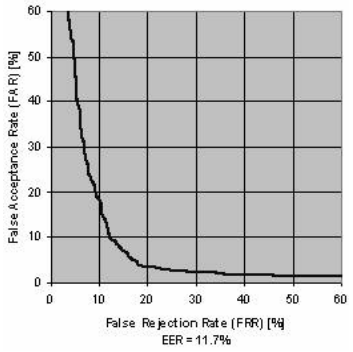
Összességben tehát az EER még alul is becsülte a különbséget ebben az esetben a két elrendezés között. Például, ha a kívánt munkaponti FRR=5%, akkor fonéma-bigramm esetén FAR=14.7%, multi-Gauss esetén 42.7%.

Az ROC görbéken végzett vizsgálataink alapján a fonéma-bigram mindentfelismerő hálózat alkalmasnak tűnik a gyakorlati alkalmazásra, bár számítási igénye kétségtelenül magasabb, mint a multi-Gauss-os változaté, de teljesítménye jóval felülmúlja azt, főként ha FRR érzékeny alkalmazásokat tekintünk.

Második típusú konfidencia-becslésünket alkalmaztuk mindkét hálózati konfigurációra, és ez újabb két görbét eredményezett. (**3. ábra**)



3.1. ábra Fonéma-bigram mindenhez illeszkedő hálózat, normált becslés



3.2. ábra Multi-Gauss mindenhez illeszkedő hálózat, normált becslés

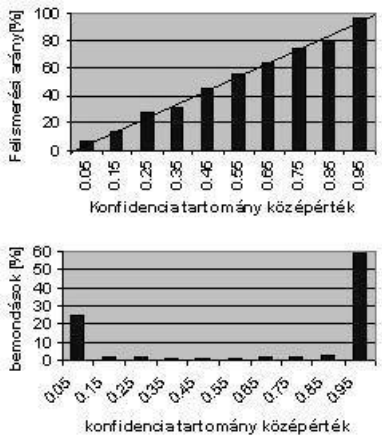
Ez az eset annyiban különbözik tehát az előzőtől, hogy a $EKB_{normált}$ érték alapján került az ROC görbe felrajzolásra. A szóhosszal való normálástól azt vártuk, hogy megbízhatóbb EKB-t szolgáltat majd, mely alapján a felismerő hatékonyabban szét tudja majd választani a szótáron belüli, illetve kívüli bemondásokat. Ezzel ellentétben mindkét esetben rosszabb EER volt mérhető. FRR=5%-os megkötésnél is romlottak a hozzá tartozó FAR értékek (15.8%, 46.7%). Kijelenthetjük tehát, hogy gyakorlati alkalmazásokban az ROC görbén végzett vizsgálat alapján nem érdemes a normált becslési módszert alkalmazni.

7. Konfidencia pontossági kiértékelés

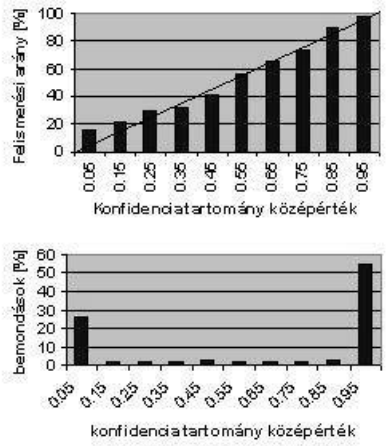
7.1. A felismerési arány grafikon

Ha a felismerés biztonságáról is szeretnénk információhoz jutni, szükség van arra is, hogy minden bemondáshoz egy hozzárendelési szabállyal (az általunk használt szabályt a cikk korábbi fejezeteiben ismertettük) egy konfidencia értéket társítsunk. Ez a fejezet a hozzárendelési szabály hatékonyságát próbálja értékelni. Ehhez szükséges bevezetni a részhalmazon mért felismerési arány fogalmát, ami azt mutatja meg, hogy a részhalmaz bemondásainak hány százalékát ismerte fel helyesen a felismerőnk. Fontos fogalom még a konfidencia tartomány, ami alatt két konfidencia érték közötti intervallumot értünk, amibe minden olyan bemondás beletartozik, amihez e két érték közötti valószínűséget rendeltünk. Minden ilyen tartományra értelmezhető egy felismerési arány is. Ha kiszámítjuk és ábrázoljuk ezt az arányt minden konfidencia tartományra, akkor jutunk a felismerési arány grafikonhoz. **[4]**

Önmagában ez az ábra még nem ad megfelelő értékelést a konfidencia-becslésről, mivel nem állapítható meg, hogy az egyes konfidencia tartományokba a szavak hány százaléka esett. Ha egy tartományban a felismerési arány jelentősen eltér a becsült konfidenciától, nem jelenti azt, hogy az egész becslés rossz volt, ha az adott tartományba a bemondásoknak elhanyagolható százaléka esett. Ezért kell kiegészíteni a felismerési arány ábrát egy grafikonnal, ami megmutatja, hogy az egyes konfidencia tartományokban a bemondások hány százaléka található (**4. ábra**).



4.1. ábra: Konfidencia-becsléshez tartozó felismerési arány grafikon, valamint a bemondások eloszlása a konfidencia függvényében. Fonéma-bigramm mindenhez illeszkedő hálózat, alapeset.



4.2. ábra: Konfidencia-becsléshez tartozó felismerési arány grafikon, valamint a bemondások eloszlása a konfidencia függvényében. Multi-Gauss mindenhez illeszkedő hálózat, alapeset.

7.2. Numerikus kiértékelés

Egy mindenhez illeszkedő hálózat felismerési arány grafikonja szemléletes, de sokszor célszerű az ábra által közölt információt egyetlen mérőszámba összevonni. Ilyen a **confidence error rate** (**CER**) százalékos mérőszám, ami megmutatja, hogy a konfidencia-becslés felismerési arány grafikonja hány százalékban tér el az ideálistól (**9**).

$$CER = \sum_{i=1}^{10} \frac{B_i}{B_f} * \left| F_i - C_i^k \right| * 100[\%] \tag{9}$$

Ahol B_f az összes bemondásnak a száma B_i pedig az i-edik konfidencia tartományba eső bemondások száma. F_i az i-edik tartomány felismerési aránya, C_i^k konfidencia tartomány középérték, pedig a tartomány alsó és felső határának számtani közepe.

7.3. Konfidencia pontosság a teljes vizsgálati halmazon

A legjobb konfidencia-becslést a korábban definiált „a” és „b” szabad paraméterek változtatásával CER minimalizálásával kerestük meg. Így kaptunk hálózatonként két konfidencia-becslést alap és normált esetre.

A két mindenhez illeszkedő hálózat CER értékei egy táblázatban kerültek összefoglalásra (**1. Táblázat**). Ezek alapján kijelenthető, hogy a fonéma-bigram mindenhez illeszkedő hálózat konfidencia-becslése hatékonyabb, ami a 4. ábrán is megfigyelhető, ha figyelembe vesszük, a 0-0.1-es konfidencia tartományt, ahol a multi-Gauss hálózat igen rosszul teljesít. Jól látható továbbá, hogy a normálás is ront konfidencia-becslés hatékonyságán.

1. Táblázat: A CER értékei az eredeti bemondáshalmaz konfidencia-becsléseire

	Fonéma-bigram	Multi-Gauss
Alap eset	2.2%	5.09%
Normált eset	3.1%	5.48%

7.4. Konfidencia-becslés pontosságának vizsgálata bemondáshalmaz szétbontással

Önmagában a teljes bemondáshalmaz vizsgálata nem elegendő, mivel az optimalizálás maga is ezen a halmazon történt. A további értékeléshez felbontottuk a bemondások adatbázisát több kritérium szerint diszjunkt bemondás-halmazokra. Először három, elemeit véletlenszerűen összeválogatott, körülbelül megegyező számú bemondást tartalmazó halmazra. A második kritérium a szótagszám szerint bontotta szét kis szavakra (1-2szótag), átlagos hosszúságú szavakra (2-4) és végül 5-nél több szótagot tartalmazó nagy szavakra. Az utolsó kritérium két csoportot hozott létre az alapján, hogy a bemondás szótáron belüli vagy kívüli.

A véletlenszerű felbontással képet kaphatunk arról, hogy az adott hálózat konfidencia-becslése (a teljes adatbázisra beállított „a" és „b" paraméterek itt már konstansok) mennyire adatbázis-független. A másik két módszerrel megfigyelhető, hogy a bemondott szó bizonyos tulajdonságai (hossza, szótáron belülisége) mennyire vannak hatással a konfidencia-becslésre.

Véletlenszerű szétbontás

A véletlenszerűen háromfelé bontott multi-Gauss és fonéma-bigram hálózatos becslések látványosan nem romlanak. A változás mértéke jó közelítéssel azonos a jellemzővektor számmal normált és nem normált becslés esetén mindkét mindenhez illeszkedő hálózat esetében. Az eredmények arra engednek következtetni, hogy a konfidencia-becslési paraméterek nem csak az optimalizációs adatbázishoz, hanem annak valamennyi részhalmazához is jól illeszkednek. (**2. Táblázat**)

2. Táblázat: A CER értéke a véletlenszerűen szétbontott bemondáshalmaz konfidencia-becslésére

	Fonéma-bigram (Alap / Normált eset)	Multi-Gauss (Alap / Normált eset)
1. Bemondáshalmaz	3.6% / 2.9%	5.7% / 4.4%
2. Bemondáshalmaz	3.3% / 3.1%	5.7% / 6.6%
3. Bemondáshalmaz	4.6% / 3.7%	6.7% / 6.3%

Szóhossz alapján történő szétbontás

A multi-Gauss hálózatos felismerőnél a szótagszám alapján történő szétbontása már változó képet mutat: a kis méretű szavak esetében a legnagyobb a változás. A legtöbb elemet tartalmazó közepes méretű szavak halmazán a romlás átlagos mértékű, a nagyméretű szavaknál javulás figyelhető meg. A normálásnak itt már volt látható hatása, kisméretű szavaknál javított a becslés hatékonyságán, 2-4 szótagos szavaknál

bár jobb nem lett, de kevésbé romlott az eredetihez képest. A fonéma-bigram típusú hálózat becslései csak a hosszú bemondásokra igazán érzékenyek. Az utóbbi csoporton egy keveset segített a normálás, sajnos annak árán, hogy az előbbi két csoport konfidencia-becslése rosszabb lett. (**3. Táblázat**)

3. Táblázat: A CER értéke a szóméret alapján szétbontott bemondáshalmaz konfidencia-becslésére

	Fonéma-bigram (Alap / Normált eset)	Multi-Gauss (Alap / Normált eset)
1-2 szótag	4% / 2.5%	6.8% / 6.9%
2-4 szótag	4.1% / 3.5%	5.8% / 5.6%
5- szótag	8.4% / 9.7%	4.5% / 3.9%

Szótári tartalom alapján történő szétbontás

A romlás minden esetben itt a leglátványosabb. A rendszer általában felismeri a szótáron belüli elemeket, néha mégis alacsony valószínűséget rendel hozzájuk. Ugyanez fordítva is igaz, több szótáron kívüli elem, túl nagy valószínűséget kap. Az ilyen jellegű hiba a nyelvi modell tulajdonságaiból ered, konfidencia-becsléssel nem javítható.

4. Táblázat: A CER értéke a szótári tartalom alapján szétbontott bemondáshalmaz konfidencia-becslésére

	Fonéma-bigram (Alap / Normált eset)	Multi-Gauss (Alap / Normált eset)
Szótáron belüli	10.2% / 9.2%	14.5% / 14.5%
Szótáron kívüli	15.5% / 15.9%	20.1% / 20.1%

8 Összefoglalás

Cikkünk folyamán egy az elméleti alapelvekre épülő megoldási lehetőséget mutattunk be a konfidencia-becslésére. Mint az látható volt, egy jellemzővektor sorozat előfordulási valószínűségének közelítése mindenhez illeszkedő hálózatok használatával célszerű és jó pontosságú becslést eredményez.

Eredményeinket áttekintve szembeötlő a két mindenhez illeszkedő hálózat alkalmazásakor jelentkező különbség. Multi-Gauss hálózat csak az ROC görbe FAR érzékeny részén volt képes a fonéma-bigram teljesítményét megközelíteni, ilyen alkalmazás azonban kevésbé gyakori. A konfidencia is jobban becsülhetőnek bizonyult a fonéma-bigram esetben, bár ez nem meglepő, mivel a konfidencia értéke EKB alapján számítódik. Ennek szótáron belüli, illetve kívüli szavak közti szeparációs képességét viszont az ROC segítségével vizsgáltuk. Az, hogy ezen vizsgálatok során multi-Gauss rosszabbul teljesített előrevetítette a rá adható pontatlanabb konfidencia-becslést. Ugyanezen szeparációs képesség hiányával magyarázható, hogy minden esetben magas CER érték adódott a bemondáshalmaz szótári tartalom alapján történő szétbontásánál. Ez mindenképpen javításra szorul a jövőben.

Fontos megjegyeznünk cikkünk egy kutatássorozat első eredményeit foglalja összse. Természetesen a konfidencia-számítási módszerünk egyszerűsége is okolható az a tapasztalható eltérésekért, ám magában rejti a továbbfejleszthetőséget, hiszen EKB-kön alkalmazott mostani lineáris transzformáció helyettesíthető magasabb szintű közelítésekkel. Sokkal meglepőbb inkább az a tény, hogy a módszer jelenlegi szintjén is már sok tekintetben pontos becslést ad.

A korábban elmondottak alapján több terület is kínál továbblépési lehetőséget. A mindenhez illeszkedő hálózatok tökéletesítése, a normált becslés átalakítása valamint összetettebb közelítések mind pontosíthatják az eljárást. Az ezek által kijelölt irányvonalak mentén kívánjunk mi is továbbfejleszteni felismerőnket. Célunk egy minden tekintetben pontosabb konfidenciát szolgáltató rendszer tervezése, ami nagy hatásfokkal képes detektálni a szótáron kívüli szavakat anélkül, hogy a szótáron belülieket elutasítaná. Egy ilyen eszköz várhatóan szükséges a valóban nyílt szótáras felismeréshez is, ami külön motiválja a kutatásainkat.

Bibliográfia

1. B. Dong, Q. Zhao, Y Yan: A fast confidence measure algorithm for continuous speech recognition (2005)
2. J. Pinto, R. N .V. Sitaram: Confidence measures in speech recognition based on probability distribution of likelihoods (2005)
3. J. Razik, O. Mella, D. Fohr, J.-P. Haton: Local word confidence measure using word graph and N-best list (2005)
4. G. Skantze: The use of speech recognition confidence scores in dialogue systems (2003)
5. Vicsi Klára et al: http://alpha.ttt.bme.hu/speech/databases.php (2005)
6. D.A.G. Williams: Knowing what you don't know: Roles for confidence measures in automatic speech recognition (1999)
7. Young, S. – Kershaw, D. – Odell, J. – Ollason, D. – Valtchev, V. – Woodland, P. The HTK Book (Version 3.0), (2000)