Új, zajbecsléssel kombinált, entrópia-alapú beszéddetektálási eljárás a beszédfelismerési hatásfok javítására

Tüske Zoltán, Mihajlik Péter, Tobler Zoltán

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék,
1117 Budapest, Magyar Tudósok körútja 2,
Hungary
tuske@alpha.tmit.bme.hu
mihajlik@tmit.bme.hu
mgen@freemail.hu

Kivonat: A küszöbszint-alapú beszéddetekció egy új változatát mutatjuk be. Az eljárás energia helyett a robusztusabb spektrális entrópiát használja a beszéd jelenlétének kijelölésére. További különlegessége és újdonsága a megközelítésnek, hogy az entrópiaszámítás előtt minimum spektrális részsáv-energiákon alapuló zajspektrum becslést használ a zaj fehérítésére. Ennek eredményeképp nagymértékben zajtűrő entrópia-alapú beszéddetekciós módszert kaptunk. Ezen állításunkat számos beszédfelismerési kísérlettel támasztjuk alá, melyekben normál és kifejezetten zajos telefonbeszéd-felismerést végeztünk. A javasolt beszéddetekciós eljárás alkalmazásával minden esetben javult a felismerési pontosság (maximálisan rel. 29,5%-kal), míg a felismerendő keretek számát nagyjából az eredeti mennyiség felére szorítva jelentősen csökkent a felismerő terhelése zajban is.

1 Bevezetés

A beszéd-alapú szolgáltatások egyre növekvő száma szükségessé teszi hatékony, zajtűrő beszéddetektorok fejlesztését. A beszéd jelenlétének kijelölése igen fontos például a beszédfelismerőknél és a beszéd telekommunikációs átvitele során.

Előbbi esetben jó beszéddetektálás esetén a felismerő csak a ténylegesen aktív szakaszokat kapja meg, a felismerő kikapcsol, ha a beszélő hallgat. A felismerés pontosabbá válhat, mert ilyenkor a nem-beszédet – amire általában a felismerő nem, vagy csak korlátozott mértékben lehet felkészült – a rendszer nem próbálja a betanított szavak valamelyikéhez hasonlítani, ezáltal a felismerő hatásfoka javul, ráadásul a számításigény is csökken. Tehát egy jó beszéddetektor képes a beszédfelismerő rendszerek pontosságán és működési sebességén javítani.

A második esetben, a beszédátvitel során, a beszéddetektálás fontosságát az adja, hogy sávszélességet spórolhatunk meg, ha a csatornán nem visszük át azokat a szakaszokat amikor a beszélő hallgat.

A távközlésben használt beszéddetektálási algoritmusok azonban nem használhatók közvetlenül a beszédfelismerésben, mert elsősorban nem a beszéd, hanem inkább a csend kijelölése a feladatuk, így nem szűrik ki a beszédfelismerést zavaró zajokat.

Az elmúlt évek során számos detektálási algoritmust dolgoztak ki a beszédfelismerés számára. Ezek az eljárások többé-kevésbé két kategóriába sorolhatók [1]. Az első típusú algoritmus ún. küszöb-alapú [1],[2],[8],[10]. Ebben az esetben a bejövő jelből beszéd/nem-beszéd eldöntésére alkalmas paraméterek kinyerése után adaptív, az idővel változó, a környezethez alkalmazkodni próbáló, vagy globális, előre beállított küszöbérték szerint történik a detektálás.

A küszöb-alapú beszéddetektálás legfontosabb lépései a következők:

- *Paraméter kinyerés*: olyan jellemzők előállítása a jelből, ami mást mutat a zaj- és mást a beszédszakaszokon.
- *Küszöbszint beállítás*: Ez alapján ítélhető meg egy jelszakaszról, hogy azt beszédnek vagy szünetnek tekintsük. Lehet adaptív vagy állandó is.

A másik típusú szegmentálási módszerek mintaillesztéses megközelítést [4] használnak. Ez esetben a beszéd mellett a zajról is szükséges modellt alkotni, és ennek paramétereit megbecsülni. A detektálás hasonlóan történik, mint a felismerési folyamat. A küszöbmódszert alkalmazó detektorokkal összehasonlítva, a mintaillesztésen alapuló eljárások tanító adatokat és nagyobb erőforrásokat igényelnek.

A továbbiakban a küszöb alapján döntő detektorokról lesz szó. Alapvetően egyszerűbbek és gyorsabbak, és jóval szélesebb az alkalmazási körük. Bár a dolgozatban elsősorban a beszédfelismerés hatásfokának javítását célozzuk a zajrezisztens beszéddetekcióval, a lehetséges alkalmazások túlmutatanak a beszédfelismerésen.

2 Energia és entrópia

2.1 Energia-alapú detektorok

Előnyük, hogy a zaj karakterisztikáját nem kell ismerni, viszont érzékenyek a nagy energiájú zajokra, hiszen nem minden beszéd, aminek energiája van, azaz jelentősen csökkenhet a detekció hatékonysága. Alacsony jel-zaj viszony (SNR = Signal to Noise Ratio) esetén pedig a halk beszédszakaszok energiáját teljesen elfedheti a zaj energiája. Tehát az energia-alapú algoritmusok rossz eredményeket mutatnak zajos körülmények között. Az aktuális, T minta hosszú t_0. keretben az energiát a következő módon számoljuk:

$$E_{jel}(t_0) = \sum_{t=t_0 \cdot T}^{t_0 \cdot T+T-1} y^2(t) \qquad (1)$$

A küszöbszint beállítása többféle módon lehetséges. Csúszó ablakos energiaátlagolással, esetleg a t_0-t megelőző rövid időintervallumból a minimális energiaszintet választva. Beszédnek pedig azokat a szakaszokat tekinthetjük, amelyek energiája a küszöb fölé – pl.: min. *6* dB-lel – emelkednek. A fentebb vázolt esetben nincs szükség spektrumszámolásra, aminek számottevő az erőforrás igénye. Bár létezik a spekt-

rum alapján számolt energia-alapú detektálás is, a spektrumból más paraméterek is kinyerhetők, és használhatók az energia mellett illetve helyett.

2.2 Spektrális entrópia-alapú beszéddetektor

E jellemző kiszámolásához szükség van a jel spektrumára. A beérkező jelet átlapolódó blokkokra bontva, és e blokkokon FFT-t (Fast Fourier Transform) végrehajtva kapjuk a jel gördülő spektrumát:

$$Y_{jel}(f,t_0)=\sum_{t=0}^{T-1} y(t_0+t)\cdot h(t)\cdot e^{-\frac{j2\pi\cdot t\cdot f}{T}} \qquad \textbf{(2)}$$

Ahol:

t: a diszkrét idő
y(t): a vizsgált jel
f: frekvencia
t_0: az aktuális keret kezedete
h(t): a súlyozó ablak (általában Hanning)

Amíg a jel-zaj viszony elég magas, addig az energia-alapú detektálás jól használható, de *SNR<0* dB esetén az eredmények elég rosszak, noha a spektrumban még jól látszanak a beszédszakaszok, a spektrum még mutat bizonyos rendezettséget. A spektrum rendezettségének mérésére, az információelméletből ismert Shannon-i entrópia mintájára, [10] bevezeti az amplitúdó spektrum entrópiáját. Ezt a következőképpen definiálja.

Az információ-forrás entrópiája (Shannon) [8]:

$$H(S)=-\sum_{i=1}^{N} P(s_i)\cdot ld\{P(s_i)\} \qquad \textbf{(3)}$$

Ahol s_i a forrásból érkező *i.* szimbólum, $P(s_i)$ az *i.* szimbólum adási valószínűsége. Ezek alapján az *t.* keret *F* frekvencián kiszámolt spektrumának entrópiája [10]:

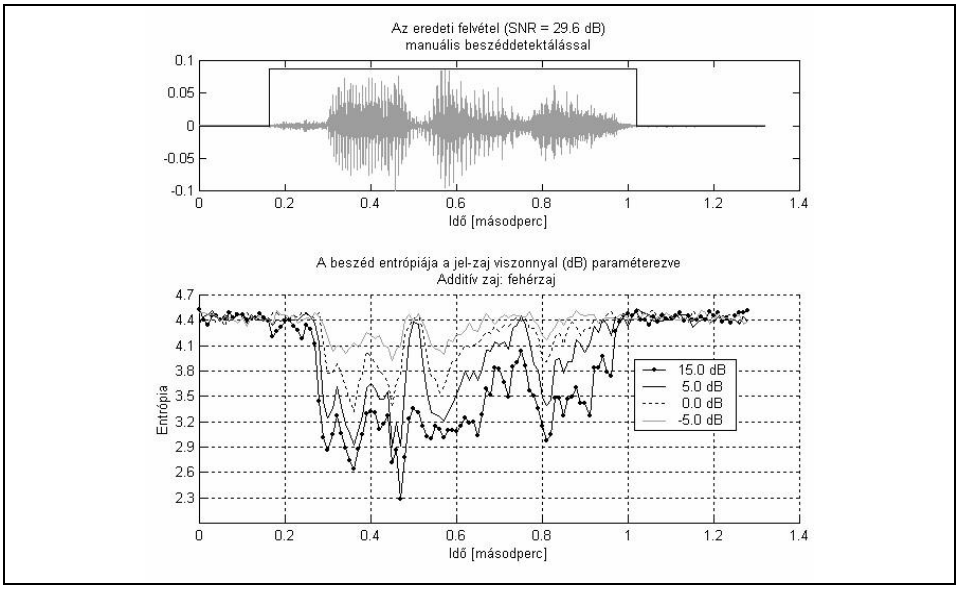
$$H\left(\left|Y_{jel}(f,t)\right|^2\right)=-\sum_{f=1}^{F} P\left(\left|Y_{jel}(f,t)\right|^2\right)\cdot ld\left\{P\left(\left|Y_{jel}(f,t)\right|^2\right)\right\} \qquad \textbf{(4)}$$

Ahol:

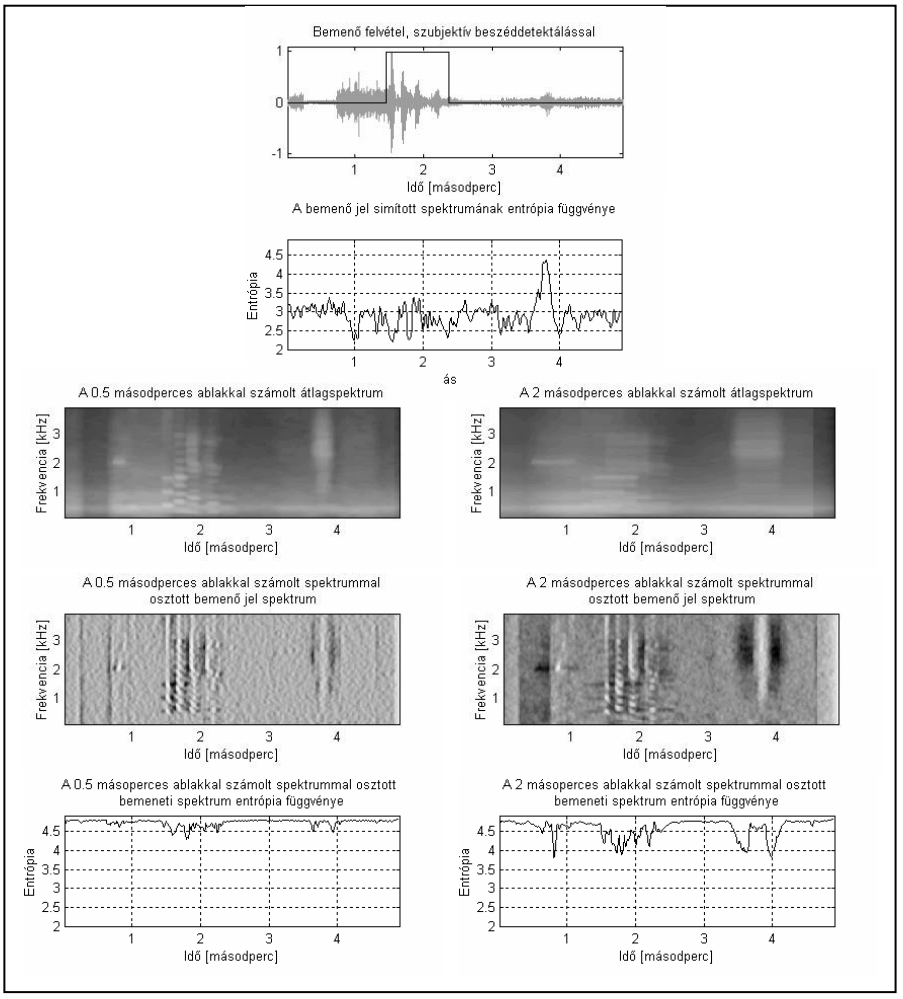
$$P\left(\left|Y_{jel}(f,t)\right|^2\right) = \frac{\left|Y_{jel}(f,t)\right|^2}{\sum_{f=1}^{F}\left|Y_{jel}(f,t)\right|^2} \quad \textbf{(5)}$$

Az entrópia egy véletlen változó bizonytalanságát írja le. Mivel a beszéd és a zaj más-más spektrális karakterisztikával rendelkezik, alkalmas paraméterválasztásnak tűnik a beszéddetektálás döntési kritériumához.

Az entrópia akkor maximális, ha a vizsgált jel fehérzaj, $H_{max} = ld(F)$; minimális, ha a jel tiszta szinusz, $H_{min} = 0$. Fontos, hogy az entrópia értéke a jelszinttől független. Így változó szintű, de állandó spektrális karakterisztikájú zaj esetén a beszéd az entrópiából könnyen kijelölhető. A küszöb meghatározható adaptívan, de létezik statisztikusan becsült megoldás is [10]. Természetesen, ha növeljük a zajszintet, akkor a beszédkeretre számolt entrópia is változik, a zaj spektruma fokozatosan elnyomja a beszédét, a spektrum végül teljesen egyenletessé válik és nem mutat rendezettséget (*1. ábra*).



1. ábra: Beszédjel entrópiájának alakulása növekvő fehérzajban



2. ábra: Az entrópia alakulása átlagspektrummal való osztás hatására

A fent leírt módszer jól használható beszéddetektáláshoz, ha a zaj fehér, azaz a spektruma egyenletes. Színes zaj esetén a zaj spektruma is rendezettebb, ezért nem lesz olyan egyértelmű a beszéd jelenléte az entrópia-idő diagramon.

A [10] az entrópia-alapú detekció egyéb zajokra való kiterjesztéséhez a következőt javasolja. Az aktuális keret spektrumát az entrópia számolása előtt osszunk le a *T* idő alatt számolt átlagoltspektrummal:

$$Y_{átlag}\left(f,t_0\right)=\frac{Y(f,t_0)}{\frac{1}{T}\sum_{t=-T/2}^{T/2}Y(f,t)} \qquad \textbf{(6)}$$

Az így kifehérített spektrumra számoljuk ki az entrópiát kiszámoljuk, és a fehérzajnál alkalmazott detektálási módszer ebben az esetben is használhatóvá válik.

Tapasztalatunk szerint a beszédszakasz spektrumát a körülötte számolt átlagspektrummal osztva lerontjuk a beszéd entrópiáját is. Tehát a zaj spektruma valóban kifehéredik, de a beszéd spektruma is. Így a fehérzajnál alkalmazott detektálási módszer nem lesz elég eredményes színes zaj esetén. (2. *ábra*)

A fenti eljárással az a probléma, hogy az átlagspektrum mindig tartalmazza a beszédspektrumot is, így az azzal való osztás mindig fehérítést jelent a beszédszakasz számára.

Természetesen adódik, hogy ha ismerjük a zaj – legalább közelítő – spektrumát, és a (6) nevezőjében az átlagspektrum helyett alkalmazzuk, akkor csak a zajspektrum fehéredik ki. Meglehet, hogy a beszédspektrum torzul ilyenkor, azonban a rendezettsége megmarad, így az entrópiája is alacsony marad, ugyanakkor a nem-beszéd szakaszok entrópiája közel maximális lesz. Ehhez tehát szükség van a beszéd alatti zaj spektrumának becslésére.

3 Zajbecslés

[7] utal egy olyan fajta zajbecslésre, ami az időben visszatekintve minden frekvenciakomponensnek a minimumát ragadja ki. Az alapgondolat, hogy a beszéd gyorsan ingadozik, szünetekkel tagolt, így megfelelően nagy T időintervallumban a frekvenciakomponensek minimumát kigyűjtve csak a zajra jellemző spektrumot kapunk, ha a zajt lassabban változónak tekintjük a beszédhez képest. A t_0 időponthoz tartozó becsült zaj spektrumát a következő módon kapjuk:

$$Y_{zaj}(f,t_0) = \min_{t=t_0-T\ldots t_0} \{Y_{jel}(f,t)\} \qquad \textbf{(7)}$$

Azonban könnyen belátható, hogy az újonnan belépő zajokkal szemben az eljárás tehetetlen, ezért az általunk javasolt zajbecslés nem csak a múltból, hanem a „jövőből" is vesz mintát a zajspektrum számításához. Természetesen a jövőbeni keretek spektrumának kiszámítása, és felhasználása csak késleltetés árán történhet meg.

A becslés hatásosságának növelésére a becsléshez használt időintervallumot két részre bontottuk, T_1 ill. T_2 hosszú szakaszokra. Mindegyikben külön-külön történt a zajbecslés, azaz két zajbecslővel. Majd a két becsült zajspektrum frekvenciakomponensei közül mindig a nagyobbikat választva került meghatározásra az aktuális keretre vonatkozó zaj spektruma. A becsült zaj t_0 idő pillanatban tehát a következő:

$$\hat{Y}_{zaj}(f,t_0) = MAX\left[\min_{t=t_0-T_1\ldots t_0}\{Y_{jel}(f,t)\}, \min_{t=t_0\ldots t_0+T_2}\{Y_{jel}(f,t)\}\right] \qquad \textbf{(8)}$$

A T_1 és T_2 értékét akkorára érdemes választani, hogy a minimumot kereső ablakban bekövetkezzen beszédhangváltozás, az amplitúdóspektrum átrendeződése. Például egy felpattanó zárhang előtt valószínűleg minden frekvenciakomponens minimumot fog elérni. A múltban működő zajbecsléshez a hosszabb időintervallumot érdemesebb használni, mint a jövő mintáiból való zajbecsléshez, mert ez nem okozhat késleltetést. Viszont a jövőből hosszabb szakaszt venni csak akkor érdemes, ha az algoritmus adatbázison fut, mert valósidejű alkalmazásoknál megengedhetetlenül nagy késleltetést vihetünk be a rendszerbe, ha túl nagy az előretekintés.

4 A javasolt detekciós algoritmus

A bemutatandó beszéddetektor algoritmust NSSE-VAD-nak neveztük (Noise-Suppressed Spectral Entropy-based Voice Activity Detection, [12]), és a következő lépéskből áll. (Lásd még:*3. ábra*)

4.1 Gördülőspektrum-számítás

A bejövő jelet *30* ezredmásodperces keretekre bontva és Hanning ablakot használva, illetve *10* ezredmásodpercenként (a keretek 66.6% átlapolódása) végzett Fourier-transzformálással számoltuk a spektrumot. Az összes beszédminta *fs = 8000* Hz –cel mintavételezett.

4.2 Simítás

Frekvenciában simított spektrumon pontosabban végezhető a zajbecslés, jobban tükrözi a sztohasztikus jelek spektrumát. Például a fehérzaj spektruma ablakozás és Fourier-transzfromálás után nem konstans, míg simítás után jobban közelíti azt. A beszéddetektálást segíti, ha az entrópia görbe gyors időbeli ingadozásait kompenzálandó időben simítjuk a gördülő spektrumot. A két műveletet elvégezéséhez, az amplitudóspektrumot idő-frekvencia síkon egyszerre simítjuk. Ehhez az alábbi 2 dimenziós FIR szűrőt, *S* mátrixot használjuk:

$$S = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 2 & 3 & 2 & 1 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \cdot \frac{1}{35} \qquad \textbf{(9)}$$

$$Y_{simitott}(f_0, t_0) = \sum_{f=-2}^{2} \sum_{t=-2}^{2} Y_{jel}(f_0 + f, t_0 + t) \cdot S(f + 3, t + 3) \qquad \textbf{(10)}$$

Zajbecslés

A zajbecslés a [7] által javasolt elgondolás továbfejlesztett változata (8) alapján történt, hogy a zajbecslő késés nélkül legyen képes követni a hirtelen belépő zajokat. A becsült zaj spektruma a minimum módszerből eredően nem lehet nagyobb egyik frekvencia-komponensen sem, mint az aktuális keret spektruma. A múltbeli zajbecslést a kísérleti tapasztalatok alapján $T_2 = 0.75$ másodpercre, a jövőből becslést pedig $T_1 = 0.25$ másodpercre választottuk.

Zajelnyomás

Az aktuális keret spektrumát (11) alapján fehérítjük. A jelspektrumból azért nem kivonjuk a zajt, mert úgy a maradékspektrum nem lenne fehér, hiszen a becsült zaj csak kisebb lehet, mint a tényleges zaj. Ugyanakkor a becsült zaj spektrumával való osztás után közel konstanssá válik a maradékspektrumban a zaj, ha jó a zajbecslés, és a zaj szerkezetét sikerül megfelelően kinyerni. Tehát az entrópia a maximálishoz közeli lesz a beszédet nem, csak zajt tartalmazó keret esetén.

$$Y_{\text{zajelnyomott}} = \frac{Y_{simitott}}{\hat{Y}_{zaj}} \qquad \textbf{(11)}$$

Spektrális entrópia számítás

Az aktuális, becsült zajjal kifehérített keret spektrális rendezettségét $H(|Y_{\text{zajenyomott}}(f,t)|^2)$ -t a (3), (4) képletek segítségével számoljuk.

Elsőszintű döntés entrópiaküszöb alapján

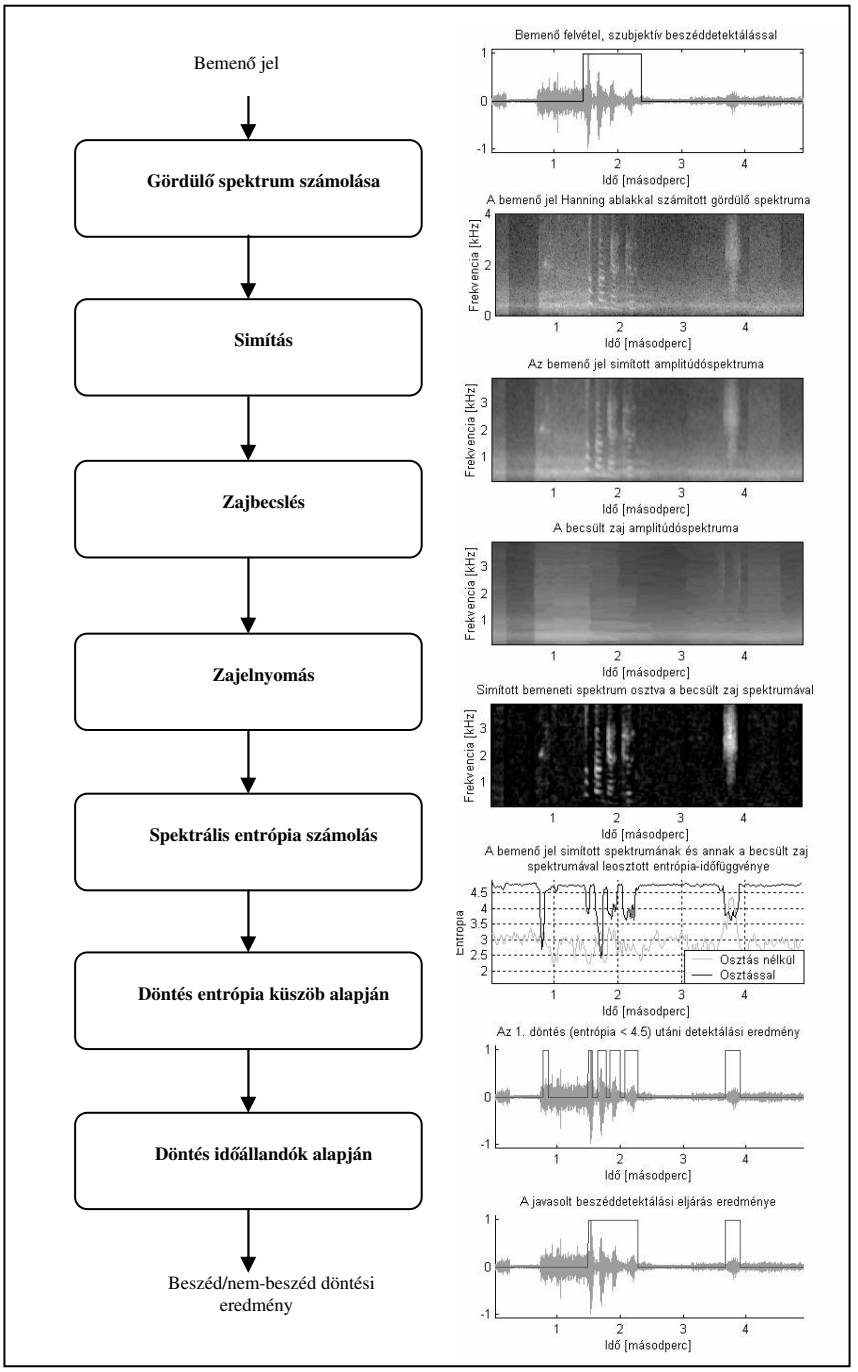
Az entrópia döntési küszöbét 4.5 –nek választottuk. E felett zajnak, alatta beszédnek tekinti a detektor az aktuális keret. Fontos hangsúlyozni, hogy ez a fajta detektálási módszer globális küszöbön alapul. Nincs szükség adaptivitásra, ez a szerep a zajbecslőé. A küszöböt empirikus módszerekkel határoztuk meg.

Második szintű döntés időállandók alapján

A beszédszakasz kijelöléséről az entrópiagörbe küszöb alá kerülésén kívül egy második réteg is dönt a következők szerint.

A beszédszakasz minimális hossza 0.2 másodperc, az ennél rövidebb beszédtartományok nem kerülnek detektálásra.

A beszédben levő szünetek áthidalására a 0.1 másodpercnél kisebb időkülönbséggel rendelkező beszédszakaszok folyamatos szakaszként kerülnek kijelölésre.



3. ábra: A javasolt detektor blokkvázlata és működése

5 Kiértékelés

A szemléltetésnél használt és számos egyéb más beszédmintán végzett kísérletek eredményei jó okot adtak arra, hogy beszédfelismerő rendszerben alkalmazva is megvizsgáljuk a detektor működését, hatását a beszédfelismerésre.

A beszéddetekció hatékonyságát indirekt vizsgáltuk. A tanszéken alkalmazott, nyilvánosan is hozzáférhető beszédadatbázissal [5] betanított beszédfelismerő rendszer felismerési hibaarányait mértük különféle lényegkiemelési beállítások mellett.

5.1 Adatbázisok

Tanításra az MTBA (Magyar nyelvű TelefonBeszéd-Adatbázis) [5] kézzel szegmentált részét használtuk. A teszteléshez két másik telefonbeszéd-adatbázist vettünk igénybe. Elsőként az MTBA-hoz nagyban hasonló Besztel adatbázis „***tiszta***”, vagyis az annotáció során nem zajosként jelölt mintegy 6000 bemondását használtuk. A másik tesztadatbázisunk a nyilvánosan is hozzáférhető Tesztel [6], „***zajos***” telefonbeszéd adatbázis volt. Az ebben levő felvételek szándékosan zajos környezetben (kocsiban, bevásárló központban, utcán, stb.), kifejezetten a zajtűrő beszédfelismerés vizsgálata végett készültek. Itt mintegy 1200 felvételt használtunk a tesztelésnél.

5.2 Vizsgálati módszer

Minden esetben 3 állapotú „balról-jobbra” struktúrájú környezetfüggő rejtett Markov-modelleket használtunk hangmodellekként. Mindkét tesztadatbázison parancsszó felismerést hajtottunk végre a „tiszta” tesztadatbázison 1000 körüli szótárérettel, míg a „zajos” adatbázison 250 körüli szótárérettel a [11] felismerővel.

Az azonos beállítású teszteket mindig párhuzamosan végeztük a két adatbázison. Ezen felül, tekintettel arra, hogy a zajos adatbázis felvételeinek jelentős része AGC (Automatic Gain Control)-torzított, minden beállításnál statikus energiával és anélkül is – az említett hatást kiküszöbölendő – elvégeztük a kísérleteket. Így tehát minden lényegkiemelési módszer esetén négy felismerési tesztet futtattunk. Végül nemcsak a javasolt detektort, hanem az ADSR (Advanced Ditributed Speech Recognition) ETSI szabványban rögzített detekciós eljárást is megvizsgáltuk.

5.3 Lényegkiemelési eljárások

A következő lényegkiemelési konfigurációk mellett végeztünk kísérleteket:

- Alkalmazva az ETSI ADSR lényegkiemelési szabványt, az abban foglalt jelalakformálást, zajelnyomást, vak csatornakiegyenlítést. (ADSR)
- Csak a Mel-frekvenciás kepsztrális együtthatókat számítva. (CC)
- A fenti mellett vak csatornakiegyenlítést is alkalmazva. (CC+BEQ)
- Csatornakiegyenlítést csak a teszteléskor végezve. (CC+fél BEQ)

5.4 Beszédfelismerési eredmények

Először beszéddetekció nélkül mértük az egyes konfigurációk hatásfokát.

1. táblázat: Referencia konfigurációk szó hibaaránya (WER = Word Error Rate) beszéddetektálás nélkül zajos és tiszta adatbázison

Lényegkiemelő	Energiával		Energia nélkül	
	Tiszta	Zajos	Tiszta	Zajos
ADSR	**5,23**	**51,24**	**6,26**	**21,20**
CC	4,78	45,61	5,26	27,33
CC+BEQ	4,76	43,60	5,43	19,97
CC + fél BEQ	4,38	41,87	4,71	20,63

Látható a referenciatáblázatban, hogy a statikus energia elhagyása igen jótékonyan hat a beszédfelismerés hatásfokára zajos esetben. Ez az AGC negatív hatásának kiküszöbölése miatt történhet. Ugyanakkor a tiszta adatokon csökken a hatásfok.

A következő mérési sorozatban a javasolt NSSE-detektor által okozott hatást vizsgáltuk a beszédfelismerés szempontjából, valamint az eredményeket az ADSR saját beszéddetekciós eljárásának erdeményeivel is összevetettük.

2. táblázat: A konfigurációk szó hibaaránya (WER, %) beszéddetektorokkal

Detektor	Lényegki-emelő	Energiával		Energia nélkül	
		Tiszta	Zajos	Tiszta	Zajos
ADSR	ADSR	5,21	51,07	6,26	21,20
NSSE	ADSR	5,11	36,14	5,86	20,54
NSSE	CC	4,66	35,51	5,08	22,77
NSSE	CC + BEQ	4,70	33,83	5,23	18,65
NSSE	CC + fél BEQ	4,27	30,94	4,51	18,48

3. tábázat: A beszéddetektor által okozott relatív százalékos javulás

Detektor	Lényegki-emelő	Energiával			Energia nélkül		
		Tiszta	Zajos	Átlag	Tiszta	Zajos	Átlag
ADSR	ADSR	+0,38	+0,33	**+0,36**	0,00	0,00	**0,00**
NSSE	ADSR	+2,29	+29,47	**+15,88**	+6,39	+3,11	**+4,75**
NSSE	CC	+2,51	+22,14	**+12,33**	+3,42	+16,68	**+10,05**
NSSE	CC + BEQ	+1,26	+22,41	**+11,83**	+3,68	+6,61	**+5,15**
NSSE	CC + fél BEQ	+2,51	+26,10	**+14,31**	+4,25	+10,42	**+7,33**

Látható, hogy a javasolt detekciós algoritmus minden esetben javított a felismerési arányon. Különösen az energiát is tartalmazó zajos eredmények kimagaslóak (maximálisan *29,47%*). Bár a szóhiba-arány eredmények is ígéretesek az NSSE-VAD és az ADSR-VAD összehasonlítást illetően, a két beszéddetektor közti különbség drámaian megnő, ha a „nem-beszéd” keretek eldobási arányait tekintjük (4. ábra).

4. tábázat: A beszéddetektorok által a felismerés során az összes keretből eldobott keretek aránya %-ban

Adatbázis	Detektor	Vektorok száma	Keret dobási arány
Tiszta	ADSR VAD	1.788.101	24,9 %
	NSSE-VAD		60,0 %
Zajos	ADSR VAD	466.332	3,5 %
	NSSE-VAD		52,6 %

7 Összefoglalás

A dolgozat során bemutatott detektálási algoritmus alkalmazásával egyrészt javultak a beszédfelismerési eredmények, másrészt az intenzív kereteldobás következtében jelentősen csökkent a felismerési folyamat erőforrásigénye. Ugyanakkor a zajbecslés az előretekintés miatt *0.25* másodperces késleltetést okoz, ami a valós idejű beszédalkalmazásoknál még megengedhető.

Bibliográfia

1. Abdallah, I., Montrèsor, S., and Baudry, M., “Speech signal detection in noisy environment using a local entropic criterion”, in Eurospeech, Rhodes, Greece, Sep. 1997.
2. Chuan JIA, Bo XU: An Improved Entropy-Based Endpoint Detection Algorithm, ICSLP’02, 2002, Beijing
3. ETSI standard doc., ETSI ES 202 050 v1.1.1.
4. E. Kosmides , E. Dermatas, G. Kokkinakis, “Stochastic endpoint detection in noisy speech”, SPECOM Workshop, 109-114, 1997.
5. http://alpha.ttt.bme.hu/speech/hdbMTBA.php
6. http://alpha.ttt.bme.hu/speech/hdbtesztelen.php
7. Izhak Shafran & Richar Rose: Robust Speech Detection And Segmentation For Real-Time ASR Application
8. Jialin Shen, Jeihweih Hung, Linshan Lee, “Robust entropy based endpoint detection for speech recognition in noisy environments”, International Conference on Spoken Language Processing, Sydney, 1998
9. Péter Mihajlik, Zoltán Tobler, Zoltán Tüske and Géza Gordos; Evaluation and Optimization of Noise Robust Front-End Technologies for the Automatic Recognition of Hungarian Telephone Speech, Eurospeech 2005, Lisbon
10. Philippe Renevey and Andrej Drygajlo: Entropy Based Voice Activity Detection in Very Noisy Conditions, Eurospeech 2001, Aalborgh
11. T. Fegyó et al. "Voxenter – Intelligent Voice Enabled Call Center for Hungarian", EUROSPEECH, pp. 1905-1908, 2003.
12. Zoltán Tüske, Péter Mihajlik, Zoltán Tobler and Tibor Fegyó; Robust Voice Activity Detection Based on the Entropy of Noisesuppressed Spectrum, Eurospeech 2005, Lisbon