Beszéddetekciós módszerek
vizsgálata és optimalizálása gépi beszédfelismerő rendszerekhez

TÜSKE ZOLTÁN, MIHAJLIK PÉTER, TOBLER ZOLTÁN, FEGYÓ TIBOR, TATAI PÉTER

Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék
mihajlik@tmit.bme.hu

Lektorált

Kulcsszavak: küszöbszint-alapú beszéddetekció, beszédfelismerés, spektrális entrópia, zajbecslés, VAD

A cikkben a küszöbszint-alapú beszéddetekcióhoz használható paramétereket vizsgáljuk. Először a beszéddetekció küszöbérték-érzékenységét analizáljuk egy kisebb teszthalmazon a különféle paraméterek mellett, majd az eredmények és gyakorlati megfontolások alapján választjuk ki a beszédfelismerési tesztekhez használt detekciós módszert. Az energia helyett a jóval robusztusabb spektrális entrópiát használjuk a beszéd jelenlétének kijelölésére. További különlegessége és újdonsága a megközelítésnek, hogy az entrópiaszámítás előtt spektrális részsáv-energiákon alapuló zajspektrum becslést használunk a zaj fehérítésére. Ennek eredményeképp nagymértékben zajtűrő, entrópia-alapú beszéddetekciós módszert kaphatunk. Ezen állításunkat számos beszédfelismerési kísérlettel támasztjuk alá, amelyekben normál, illetve kifejezetten zajos telefonbeszéd-felismerést végeztünk. A javasolt beszéddetekciós eljárás alkalmazásával minden esetben javult a felismerési pontosság (maximálisan 29,5%-kal), valamint a felismerendő keretek száma is jelentősen csökkent mind tiszta, mind zajos felvételek esetén.

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 kissebességű beszédátvitel során.

Az előbbi esetben, hatékony beszéddetektálás esetén, a felismerő csak a beszédet tartalmazó kereteket kapja meg, így a felismerő beszédszünetekben kikapcsolható. A felismerés pontosabbá válhat, mert ilyenkor a nem-beszédet – amit általában a felismerő nem, vagy csak korlátozott mértékben tud kezelni – 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 közismerten azért fontos, mert sávszélességet spórolhatunk meg, ha a csatornát beszédszünetekben nem foglaljuk. A távközlésben használt beszéddetektálási algoritmusok azonban közvetlenül nem használhatók a beszédfelismerésben, mert elsősorban nem a beszéd, hanem inkább a csend kijelölése a feladatuk, így nem szűrik hatékonyan a beszédfelismerést zavaró, nagyszintű 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, úgynevezett **küszöbalapú** [1,2,9,11]. 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, amelyek értéke más a zaj- és más a beszédszakaszokon.
– *Küszöbszint beállítás:*
  ennek alapján ítélhető meg egy jelszakaszról, hogy azt beszédnek vagy szünetnek tekintsük. Lehet adaptív vagy állandó is.

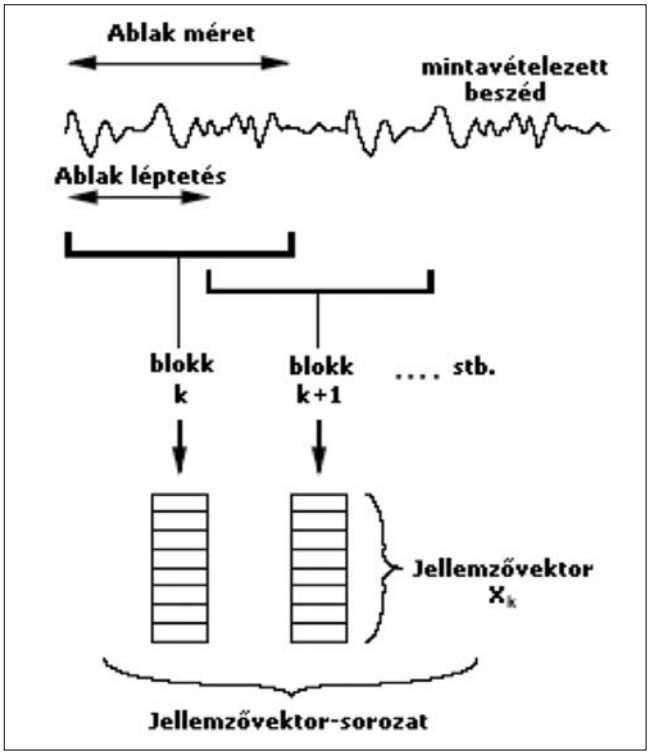
A másik elterjedt megközelítés a **mintaillesztésen alapuló beszéddetektálás** [4]. Ebben az esetben nemcsak a beszédről, hanem a zajról is modellt kell alkotni, és ennek paramétereit megbecsülni. A detektálás hasonlóan történik, mint maga a felismerési folyamat. A küszöb alapú módszert alkalmazó detektorokkal 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 túlmutatnak a beszédfelismerésen.

2. Beszéddetekciós paraméterek

A jelből olyan paramétereket célszerű kinyerni, amelyek különböző eloszlást mutatnak a beszédre és a nem-beszédre. Az egyes állapotok eloszlásának mérésére megfelelő adatbázis szükséges, az adott felvételeket pontosan fel kell címkézni.

A beérkező jel *k*. szakaszából *L* dimenziós paramétert kinyerve áll elő X_k paraméter-*oszlopvektor*. A detektálás során a paramétervektor alapján történik a



1. ábra Paramétervektor előállítása a jelből

döntés az előre felvett állapotok valamelyikére (H_i): a beszédre és a nem-beszédre. Ha az állapotok számát illetően csak kétféle osztályozás történhet (H_0, H_1), akkor a döntés a következő formában írható:

$$P(X_k|H_0)\cdot P(H_0) \underset{H_1}{\overset{H_0}{\gtrless}} P(X_k|H_1)\cdot P(H_1) \qquad (1)$$

ahol H_0: az aktuális keret nem-beszéd,
H_1: az aktuális keret beszéd.

Átrendezve és $\frac{P(H_0)}{P(H_1)}$ helyett más küszöböt, η-t választva, skálázhatóbbá válik a detektálás.

$$\frac{P(X_k|H_1)}{P(X_k|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \qquad (2)$$

Többdimenziós X_k esetén az *i.* állapothoz tartozó eloszlást általában Gauss-eloszlással közelítik. Egydimenziós paraméterek esetén könnyen mérhető és ábrázolható az (1)-es képletben szereplő, egyes állapotokra jellemző eloszlás sűrűségfüggvénye. A kétállapotú döntés miatt a küszöbérték kiindulási értékének a beszédhez és a nem-beszédhez tartozó paraméter-eloszlásfüggvények metszéspontja tekinthető. Ekkor az aktuális keretben mért paraméterérték alapján igen egyszerűen dönthetünk beszédre, illetve nem-beszédre.

2.1. Energia

Az energiaküszöb-alapú megközelítés előnye, hogy a zaj karakterisztikáját nem kell ismerni. Hátránya viszont, hogy érzékeny a nagy energiájú zajokra, hiszen nem minden beszéd, aminek nagy 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. kezdetű keretben (ahol mintavett, azaz diszkrét idejű jelet dolgozunk fel) az energiát a következő módon számoljuk:

$$E_{jel}(t_0) = \sum_{t=t_0}^{t_0+T-1} y^2(t) \qquad (3)$$

A küszöbszint beállítása többféle módon lehetséges, például csúszó ablakos energiaátlagolással, vagy a t_0-t megelőző rövid időintervallumból a minimális energiaszintet választva. Beszédnek pedig azokat a szakaszokat tekinthetjük, amelyek energiája – például min. *6* dB-lel – a küszöb fölé emelkednek. A fentebb vázolt esetben nincs szükség spektrumszámolásra, aminek számottevő az erőforrás igénye. Bár létezik a spektrum alapján számolt energia-alapú detektálás is, a spektrumból más paraméterek is kinyerhetők és használhatók az energia mellett, illetve helyette.

2.2. Spektrális entrópia

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

$$Y_{jel}(f,t_0) = \sum_{t=0}^{T-1} y(t_0+t)\cdot h(t)\cdot e^{-\frac{j2\pi\cdot t\cdot f}{T}}, \qquad (4)$$

ahol: t: a diszkrét idő,
$y(t)$: a vizsgált jel,
f: frekvencia,
t_0: az aktuális keret kezdete,
$h(t)$: a súlyozó ablak (általában Hanning).

Amíg a jel-zaj viszony elég nagy, addig az energia-alapú detektálás jól használható, de *SNR <0* dB esetén az eredmények már elég rosszak, noha a spektrumban még jól látszanak a beszédszakaszok, vagyis a spektrum még mutat bizonyos rendezettséget. A spektrum rendezettségének mérésére az információelméletből ismert Shannon-i entrópia mintájára [11] bevezeti az amplitúdó spektrum entrópiáját. Ezt az alábbiak szerint definiálja. Az információ-forrás entrópiája (Shannon) [9]:

$$H(S) = -\sum_{i=1}^{N} P(s_i)\cdot ld\{P(s_i)\}, \qquad (5)$$

ahol s_i a forrásból érkező *i.* szimbólum, $P(s_i)$ az *i.* szimbólum adási valószínűsége. Ezek alapján a *t.* keret *f* frekvencián kiszámolt spektrumának entrópiája [11]:

$$H\left(|Y_{jel}(f,t)|^2\right) = -\sum_{f=1}^{F} P\left(|Y_{jel}(f,t)|^2\right)\cdot ld\left\{P\left(|Y_{jel}(f,t)|^2\right)\right\}, \qquad (6)$$

ahol:

$$P\left(|Y_{jel}(f,t)|^2\right) = \frac{|Y_{jel}(f,t)|^2}{\sum_{f=1}^{F}|Y_{jel}(f,t)|^2}. \qquad (7)$$

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

Az entrópia maximális, ha a vizsgált jel fehérzaj, H_{max}= *ld(F)*; és 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 beszédszakaszok az entrópiából könynyen kijelölhetők. A küszöb meghatározható adaptívan, de létezik statisztikusan becsült megoldás is [11].

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 *(2. ábra)*.

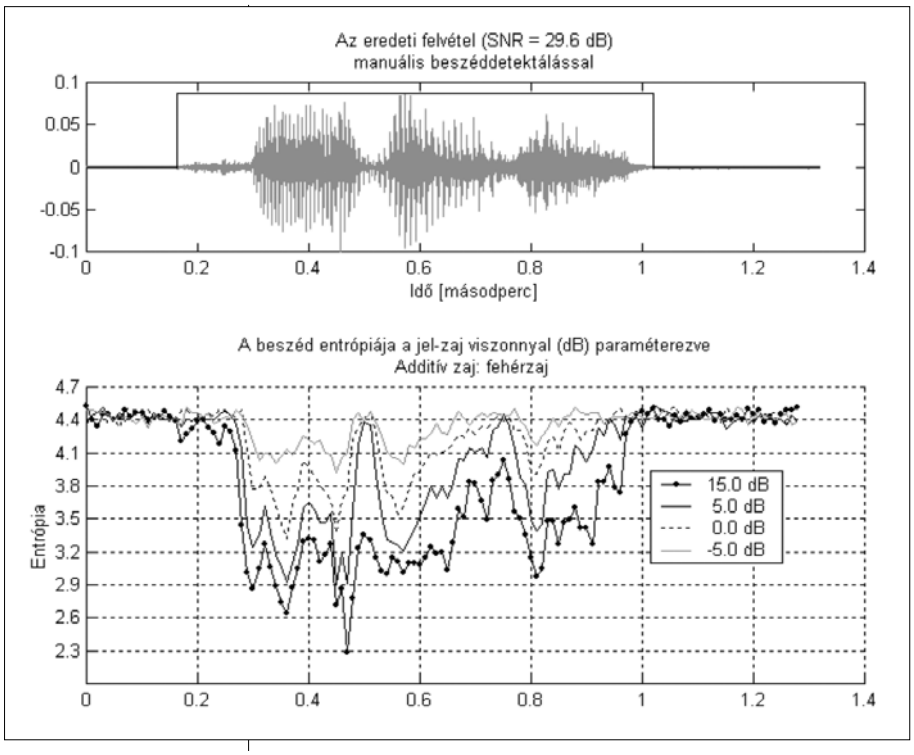
A spektrális entrópiaküszöbmódszer tehát jól használható beszéddetektáláshoz, ha a zaj fehér, azaz a spektruma egyenletes. Színes zaj esetén a zaj spektruma is rendezettebb, ezért nem lesz olyan egyértelmű a beszéd jelenléte az entrópia-idő diagramon.

A [11] irodalom az entrópia-alapú detekció egyéb zajokra való kiterjesztéséhez a következőt javasolja. Az aktuális keret spektrumát az entrópia számolása előtt osszuk el a *T* időre átlagolt spektrummal (8):

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

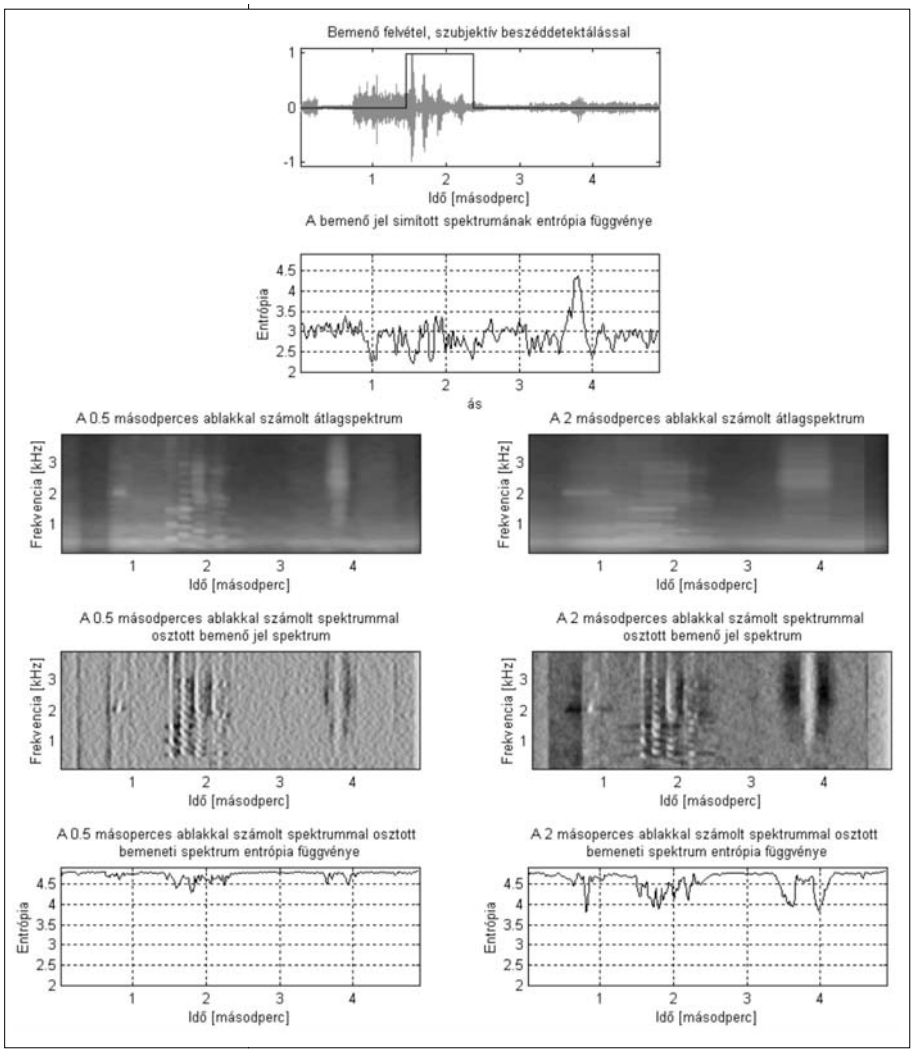
Az így kifehérített spektrumra számoljuk ki az entrópiát, és így a fehérzajnál alkalmazott detektálási módszer ebben az esetben is használhatóvá válik.

Tapasztalatunk szerint a beszédszakasz spektrumát a körülötte számolt átlagspektrummal osztva lerontjuk a beszéd entrópiáját is. Tehát a zaj spektruma



2. ábra Beszédjel entrópiájának alakulása fehérzajban

3. ábra Az entrópia alakulása átlagspektrummal való osztás hatására



valóban kifehéredik, de tulajdonképpen a beszéd spektruma is. Így a fehérzajnál alkalmazott detektálási módszer nem lesz elég eredményes színes zaj esetén *(3. 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 (8) nevezőjében az átlagspektrum helyett azt 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.

2.3. Hosszúidejű spektrális divergencia

A [8] alapján, ha ismerjük a jel gördülő amplitúdóspektrumát, $X_{k,l}$-t, ahol *k* a diszkrét időt, *l* a frekvenciasávot jelöli, akkor a jel *N*-ed rendű hosszúidejű spektrális „burkolója” (LTSE – Long-Term Spectral Envelope):

$$LTSE_N(k,l) = \max\left\{X_{k+j,l}\right\}_{j=-N}^{j=+N} \tag{9}$$

A *k*. keret hosszúidejű spektrális divergenciáját a (10) szerint kapjuk meg, az időben átlagolt zaj-amplitúdóspektrummal ($X_{Noise}(l)$) osztott *LTSE(k,l)* frekvenciakomponenseiből képzett átlagnak a logaritmusával (L jelöli a frekvenciasávok számát):

$$\mathrm{LTSD}_N(k) = 10\log_{10}\left(\frac{1}{L}\sum_{l=0}^{L-1}\frac{LTSE^2(k,l)}{X_{Noise}^2(l)}\right) \tag{10}$$

A képlet hasonló ahhoz, mintha minden frekvenciakomponensen jel-zaj viszonyt mérnénk, és átlagolnánk ezeket. A [8] állítása szerint ez a tényező egészen mást mutat zaj és mást beszéd esetén. A jó eredményhez persze szükség van a zaj spektrumának becslésére.

2.4. LPC együtthatók

A jelből kinyerhető LP (Linear Prediction) együtthatók alkalmasak a beszédspektrum burkolójának kinyerésére, a beszéd átvitele során lényegkiemelésre és tömörítésre. Az LPC együtthatók alapján történő beszédtömörítés alapja, hogy a beszéd spektrumát csak pólusokkal is jól lehet közelíteni, hiszen a zöngés hangok alap- és felharmónikus frekvenciái megfeleltethetőek az LPC-ből képzett szűrő pólusainak. Az *X(n)* jelből lineárisan predikált $X_p(n)$ jel alakja (11):

$$X_p(n) = -a_1 \cdot X(n-1) - a_2 \cdot X(n-1) - \ldots - a_N \cdot X(n-N)$$

Az a_k együtthatók meghatározása a becslés négyzetes hibájának, $\sum(X(n) - X_p(n))^2$, minimalizálásával történik. Az LP szűrő az „$X(n)$-$X_p(n)$”-t hibajelet állítja elő, és a *z* tartományban a következő módon írható:

$$H(z) = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_N z^{-N}}. \tag{12}$$

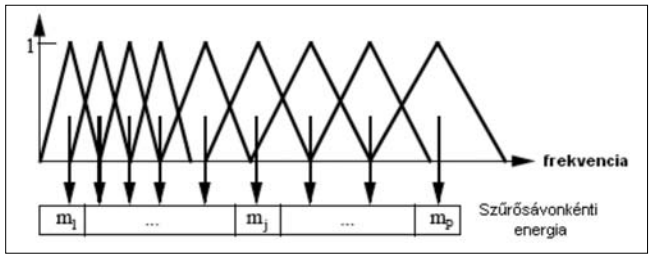
Az LP együtthatók alapján készült szűrő tehát jellemzi a beszéd spektrumát.

2.5. Mel-kepsztrum

Az amplitúdóspektrum egyenletes frekvenciaosztásokkal tartalmazza az adott keret energiájának eloszlását. Az emberi hallás azonban nem egyformán érzékeny az egyes frekvenciaközökre. Az emberi hallás frekvenciában nemlineáris karakterisztikáját figyelembe vehetjük, ha az adott keret spektrális energia-eloszlását lineáris Mel-skálán számoljuk ki. Az *f* frekvencia megfelelője a Mel-skálán [15]:

$$M(f) = 2595\log_{10}\left(1 + \frac{f}{700}\right). \tag{13}$$

A nemlineáris karakterisztika megvalósításának egyik módszere az, ha a jelet időben szűrjük Mel-skála szerint elosztott sávszűrőkkel, és a sávokra különkülön számoljuk keretenként az energiát. A másik, és a gyakorlatban inkább használt módszer, ha az aktuális jelszakaszt Fourier-transzformáljuk, majd az egyes szűrősávokra eső energiát összegezzük a megfelelően változó számú frekvenciakomponensekre.



4. ábra Melszűrőbank és a szűrőnkénti energiák

A beszédfelismerésben azonban tipikusan nem a Mel-spektrumot, hanem annak egy származtatott menynyiségét, a Mel-kepsztrumot használjuk. Ebben az esetben az együtthatókat a jel Mel-skálás reprezentációjából DCT (Discrete Cosinus Transform) használatával nyerjük. Az *i*. MFCC (Mel-Frequency Cepstral Coefficient) együttható képlete:

$$c_i = \sqrt{\frac{2}{N}}\sum_{j=1}^{N}\lg(m_j)\cos\left(\frac{\pi \cdot i}{N}(j-0.5)\right). \tag{14}$$

Itt *N* a Mel-szűrőbankok száma, m_j a *j*. Mel-szűrőn mért energia az aktuális keretben. A Mel-kepsztrum együtthatókból általában nincs szükség az összesre, csak az első *p* darabra (általában p=12).

2.6. Kepsztrális divergencia

A [12] irodalom bevezeti a kepsztrális koefficienst *V*-t, ami nem más, mint a jel kepsztrális együtthatói négyzetének összege, azaz a keretenkénti kepsztrális együtthatók második momentuma.

$$V_1 = \frac{1}{D}\sum_{i=1}^{D} c_i^2. \tag{15}$$

3. Zajbecslés

A beszéd detektálásához mindig szükség van valamilyen beszédjellemző paraméterre, amelyekről az előző fejezetben adtunk áttekintést. Azonban a zaj-rezisztens

beszéddetekcióhoz általában szükség van még a zajjellemzők (tipikusan a zajspektrum) becslésére is.

[7] utal egy olyan fajta zajbecslésre, ami az időben visszatekintve minden frekvencia-komponensnek 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 kapjuk, ha a zajt lassabban változónak tekintjük, mint a beszédet. A t_0 időponthoz tartozó becsült zaj spektrumát a következő módon kapjuk:

$$Y_{zaj}(f,t_0) = \min_{t=t_0-T\ldots t_0}\{Y_{jel}(f,t)\} \qquad (16)$$

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, illetve T_2 hosszú szakaszokra. Mindegyikben külön-külön történt a zajbecslés, azaz két zajbecslővel. Majd a két becsült zajspektrum frekvenciakomponensei közül mindig a nagyobbikat választva határoztuk meg az aktuális keretre vonatkozó zaj spektrumát. A becsült zaj t_0 időpillanatban tehát a következő (17):

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

T_1 és T_2 értékét akkorára érdemes választani, hogy a minimumot kereső ablakban bekövetkezzen beszédhangváltozás, vagyis az amplitúdóspektrum átrendeződése.

Például egy felpattanó zárhang előtt valószínűleg minden frekvenciakomponens minimumot fog elérni. A múltban működő zajbecsléshez 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.

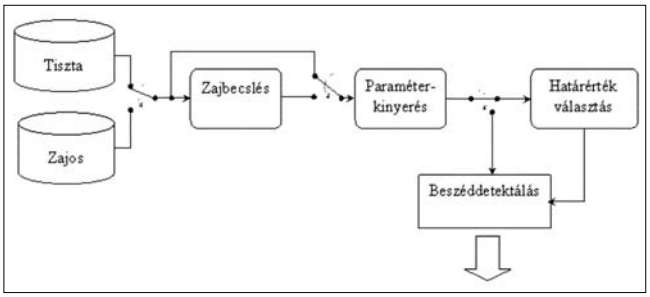
4. A beszéddetekciós paraméterek összehasonlítása ROC görbékkel

A szakirodalom a beszéddetektorokat az úgynevezett ROC (Receiver Operating Characteristics) görbékkel jellemzi, hasonlítja össze egymással. Ennek lényege, hogy a küszöbérték függvényében ábrázoljuk a detektálási eredményeket, a „mindent beszédnek detektálástól" a „mindent szünetnek detektálásig".

A grafikon x tengelyén a nem-detektált beszédszakaszok arányát (False Alarm Rate H0 = FAR0), az y tengelyen pedig a helyesen detektált szünet arányát (Hit Rate H0 = HR0) ábrázoljuk. Adott küszöb mellett ez meghatároz egy (x,y) pontot. A különféle küszöbszintekhez tartozó pontok összessége adja meg a vizsgált paramétert használó beszéddetektor ROC görbéjét. Ebben az értékelési módban nem játszik szerepet a beszéd- és a nem-beszédkeretek egymáshoz viszonyított mennyisége. Az a jobb detektor, amelyik a (0,1) ideális pontot minél jobban megközelíti, illetve amelyik ROC görbéje a nagyobb.

A 2. fejezetben említett beszéddetektálási paramétereket az *5. ábrán* látható mérési elrendezésben teszteltük.

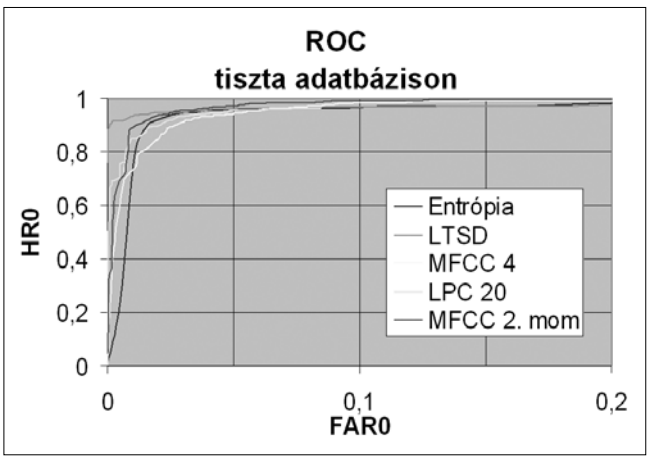
Az egyes paramétereket külön-külön optimalizáltuk. A zajos beszédanyag a [6] adatbázis 100 felvételből álló részhalmaza (mindegyik felvétel más beszélőtől származik). A tiszta adatbázis pedig az [5] adatbázis nem publikus tüköradatbázisának (Besztel) szintén 100 beszélőtől származó részhalmaza. Mindkét adatbázis valóságos (nem laboratóriumi) környezetben felvett telefonbeszédet tartalmaz, de az első esetben a beszélők kifejezetten arra lettek kérve, hogy zajos helyről telefonáljanak, míg a második esetben az utólag zajosnak minősített felvételeket nem válogattuk be a tesztalmazba.

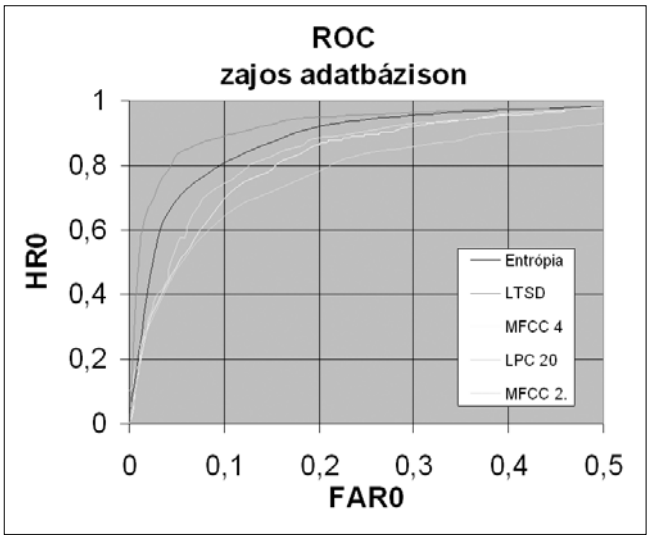


5. ábra A mérési elrendezés

A zajos és a tiszta adatbázisokon mért eredmények a *6. és a 7. ábrán* láthatóak. A mérések az entrópia, az LTSD, a 40 Mel szűrőből számolt 4 MFC együttható és az MFCC 2. momentuma típusú paraméterek esetén zajbecsléssel történtek. A 20 LP együttható pedig zajbecslés nélkül volt optimális.

6. ábra
Detektortípusok összehasonlítása tiszta adatbázison





6. ábra
Detektortípusok összehasonlítása zajos adatbázison

Mind a zajos mind a tiszta adatbázison mért ROC görbék világosan mutatják, hogy az LTSD paraméter a legmegfelelőbb a küszöbszint-alapú beszéddetekcióra a vizsgált paraméterek közül. Azonban az LTSD, vagyis a hosszú idejű spektrális divergencia számítása olyan nagy előretekintő időablakot igényel, ami az on-line rendszereknél nem engedhető meg. Így az előzetes ROC analízis második legjobban teljesítő jelöltjét, a zajbecsléssel korrigált spektrális entrópia-alapú beszéddetekciós megközelítést választottuk ki implementálásra és további beszédfelismerési vizsgálatokra.

5. A javasolt beszéddetekciós algoritmus

A bemutatandó beszéddetektor algoritmust NSSE-VAD-nak neveztük (Noise-Suppressed Spectral Entropy-based Voice Activity Detection, [14]), és a következő lépésekből áll (lásd a *8. ábrát*):

5.1. Gördülőspektrum-számítás

A bejövő jelet 30 ezredmásodperces keretekre bontva és Hanning ablakot használva, illetve 10 ezredmásodpercenként (a keretek 66,6% átlapolódásával) végzett Fourier-transzformálással számoltuk a spektrumot. Az összes beszédmintát fs = 8000 Hz-cel mintavételeztük.

5.2. 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 művelet elvégezéséhez az amplitúdóspektrumot az idő és a frekvencia síkon egyszerre simítjuk.

Ehhez az alábbi S mátrix-szal adott kétdimenziós FIR szűrőt használjuk (18):

$$S = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 2 & 3 & 2 & 1 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \cdot \frac{1}{35}$$

$$Y_{simitott}(f_0, t_0) = \sum_{f=-2}^{2} \sum_{t=-2}^{2} Y_{jel}(f_0 + f, t_0 + t) \cdot S(f+3, t+3) \quad (19)$$

5.3. Zajbecslés

A zajbecslés a [7] által javasolt elgondolás továbbfejlesztett változata volt, ami (17) alapján úgy történt, hogy a zajbecslő késés nélkül képes volt 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 jövőbeli becslést pedig T_1 = 0,25 másodpercre választottuk.

5.4. Zajelnyomás

Az aktuális keret spektrumát (20) alapján fehérítjük. A jelspektrumból azért nem kivonjuk a zajt, mert ha az aktuális keret valódi zajának spektruma nem konstans, akkor a kivonás után a maradék spektrum sem fehér lenne, hiszen a becsült zaj csak kisebb lehet, mint a tényleges zaj. Ugyanakkor az aktuális keret spektruma a becsült zaj spektrumával való osztás után közel konstanssá válik. Tehát az entrópia a maximálishoz közeli lesz olyan keret esetén, amely beszédet nem, csak zajt tartalmaz.

$$Y_{zajelnyomott} = \frac{Y_{simitott}}{\hat{Y}_{zaj}} \quad (20)$$

5.5. Spektrális entrópia számítás

Az aktuális, becsült zajjal kifehérített keret spektrális rendezettségét $H(|Y_{zajelnyomott}(f,t)|^2)$-t a (6),(7) képletek segítségével számoljuk.

5.6. Elsőszintű döntés entrópiaküszöb alapján

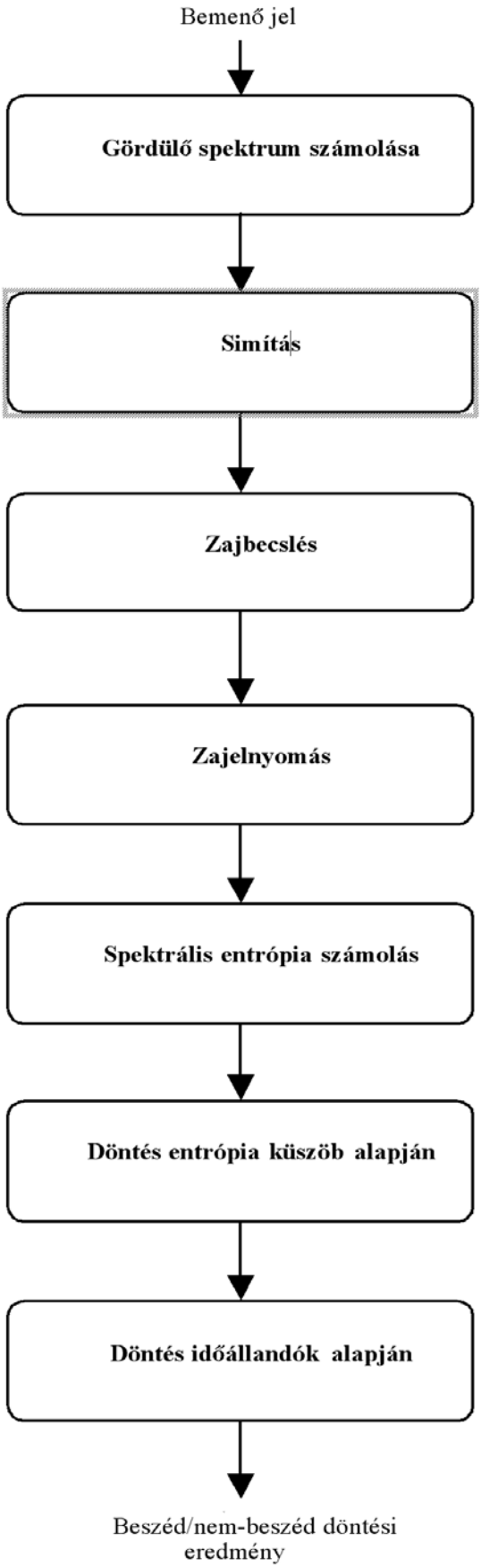
Az entrópia döntési küszöbét 4.5-nek választottuk. E felett zajnak, alatta beszédnek tekinti a detektor az aktuális keret. Fontos hangsúlyozni, hogy ez a fajta detektálási módszer globális küszöbön alapul. Nincs szükség adaptivitásra, ez a szerep a zajbecslőé. A küszöböt empirikus módszerekkel határoztuk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

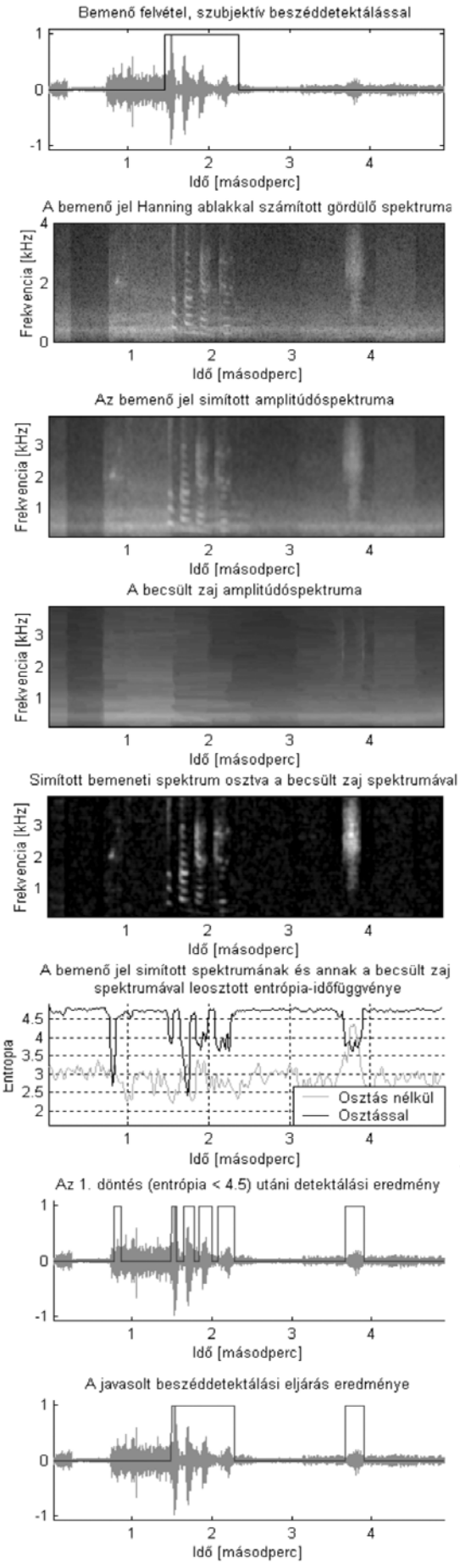
5.7. Második szintű döntés időállandók alapján

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.

8. ábra A javasolt detektor blokkvázlata és működése





6. Kiértékelés

Az ROC analízis, valamint a számos beszédfelvételen elvégzett szubjektív beszéddetekciós 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 hibaarányát mértük különféle lényegkiemelő konfigurációs beállítások mellett.

6.1. 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 természetes zajos környezetben (kocsiban, bevásárlóközpontban, utcán stb.), kifejezetten a zajtűrő beszédfelismerés vizsgálata végett készültek. Itt mintegy 1200 felvételt használtunk a tesztelésnél.

6.2. Vizsgálati módszer

Minden esetben 3 állapotú, „balról-jobbra” struktúrájú, környezetfüggő, rejtett Markov-modelleket használtunk hangmodellként. Mindkét tesztadatbázison parancsszó felismerést hajtottunk végre, a „tiszta” tesztadatbázison 1000 körüli szótárérettel, míg a „zajos” adatbázison 250 körüli szótárérettel, mindkét esetben a [13] felismerővel. Az azonos beállítású teszteket mindig párhuzamosan végeztük a két adatbázison. 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 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.

6.3. 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)

6.4. Beszédfelismerési eredmények

Először beszéddetekció nélkül mértük az egyes konfigurációk hatásfokát.

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

1. táblázat Referencia konfigurációk szó hibaaránya (WER – Word Error Rate, %) beszéddetektálás nélkül, zajos és tiszta adatbázison

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 kissé csökken a hatásfok.

Detektor	Lényeg-kiemelő	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

Detektor	Lényeg-kiemelő	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**

2. és 3. táblázat
A konfigurációk szó hibaaránya beszéddetektorokkal
A beszéddetektor által okozott relatív százalékos javulás

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

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 %

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

7. Összefoglalás

Többféle, zajtűrő beszéddetektáláshoz használatos paramétert vizsgáltunk meg. A ROC analízis alapján a praktikusan megvalósítható spektrális entrópia-küszöbön alapuló beszéddetekciós módszert választottuk ki implementálásra az általunk javasolt zajbecsléssel kiegészítve.

Megközelítésünket összevetettük az ETSI ADSR szabványában rögzített beszéddetekciós módszerrel. Az általunk használt, természetes háttérzajjal terhelt és háttérzaj-mentes telefonbeszédadatbázisokon a bemutatott detektálási algoritmus alkalmazásával egyrészt javultak a beszédfelismerési eredmények, másrészt az intenzív kereteldobás következtében jelentősen csökkent a felismerési folyamat erőforrásigénye. A zajbecslés az előretekintés miatt 0,25 másodperces késleltetést okoz, ami a valós idejű beszédalkalmazásoknál még megengedhető.

Irodalom

[1] Abdallah, I., Montrèsor, S., Baudry, M., „Speech signal detection in noisy environment using a local entropic criterion”, in Eurospeech, Rhodes, Greece, September 1997.

[2] Chuan JIA, Bo XU: Improved Entropy-Based Endpoint Detection Algorithm, ICSLP’02, Beijing, 2002.

[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 1997., pp.109–114.

[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, Proc. of IEEE Int’l Conf. on Acoustic Signal and Speech Processing (ICASSP), Hong Kong, 2003. Vol.1, pp.432–445.

[8] Javier Ramírez, José C. Segura, Carmen Benítez, Ángel de la Torre, Antonio Rubio, „Efficient voice activity detection algorithms using long-term Speech information”, Speech Communication 42 (2004), pp.271–287.

[9] Jialin Shen, Jeihweih Hung, Linshan Lee, „Robust entropy based endpoint detection for speech recognition in noisy environments”, International Conf. on Spoken Language Processing, Sydney, 1998.

[10] Péter Mihajlik, Zoltán Tobler, Zoltán Tüske, Géza Gordos; Evaluation and Optimization of Noise Robust Front-End Technologies for the Automatic Recognition of Hungarian Telephone Speech, Eurospeech 2005, Lisbon.

[11] Philippe Renevey, Andrej Drygajlo: Entropy Based Voice Activity Detection in Very Noisy Conditions, Eurospeech 2001, Aalborgh.

[12] Sergei Skorik, Frédéric Berthommier, „On a cepstrum-based speech detector robust to white noise”, SPECOM Workshop, St. Petersburg, 2000.

[13] T. Fegyó et al. „Voxenter – Intelligent Voice Enabled Call Center for Hungarian”, EUROSPEECH 2003. pp.1905–1908.

[14] Zoltán Tüske, Péter Mihajlik, Zoltán Tobler, Tibor Fegyó; Robust Voice Activity Detection Based on the Entropy of Noisesuppressed Spectrum, Eurospeech 2005, Lisbon.

[15] Steve Young at al.: The HTK Book, Cambridge, 2001.