HANGVEZÉRELT ESZKÖZÖK IPARI KÖRNYEZETBEN

VOICE CONTROLLED DEVICES IN INDUSTRIAL ENVIRONMENT

Czap László[*]*, Pintér Judit*[**]

ABSTRACT

The most comfortable way of human communication is speech, which is a possible channel of human-machine interface as well. Moreover, a voice driven system can be controlled with busy hands. Performance of a speech recognition system is highly decayed by presence of noise. Logistic systems typically work in noisy environment, so noise reduction is crucial in industrial speech processing systems. Traditional noise reduction procedures (e.g. Wiener and Kalman filters) are effective on stationary or Gaussian noise. The noise of a real workplace can be captured by an additional microphone: The voice microphone takes both speech and noise, while the noise mike takes only the noise signal. Because of the phase shift of the two signals, simple subtraction in time domain is ineffective. In this paper, we discuss a spectral representation modeling the noise and voice signals. A frequency spectrum based noise cancellation method is proposed and verified in real industrial environment.

1. BEVEZETÉS

A beszéd az emberek közötti legtermészetesebb és leggyorsabb magas szintű kommunikációs forma. Az ember régi vágya, hogy az általa konstruált gépekkel, berendezésekkel emberi nyelven, a beszéd eszközével tudjon hatékonyan és megbízhatóan kommunikálni, hasonló tempóban, mintha két ember beszélgetne egymással. A beszéd alapú kommunikáció bizonyos esetekben egyéb előnyökkel is jár. Például logisztikai rendszereket úgy is vezérelhetünk, ha mindkét kezünk foglalt.

Azok a beszédfelismerők, amelyek laboratóriumi környezetben rögzített hangmintákkal lettek betanítva, ipari zajos környezetben használhatatlanná válnak. A laboratóriumi beszédfelismerők megfelelően működnek saját környezetükben, ahol a beszédjelet anélkül lehet rögzíteni, hogy azt zaj terhelné, vagy hallható lenne másik beszélő hangja, vagy más zavaró jel. Ahhoz, hogy a beszédfelismerők ipari zajos környezetben is megfelelő hatékonysággal működjenek, a következő lépés a rendszer módosítása.

A hagyományos beszédfelismerő módszerek egy mikrofonosak. Ezeken a módszereken alapuló beszédfelismerők megfelelően működnek alacsony zajszintű környezetben, amit egyedi mikrofonnal terveztek vagy tanítottak. Amikor azonban a környezet zajjal terhelt, vagy interferáló jeleket tartalmaz, vagy ha eltérő karakterisztikájú mikrofonnal rögzítjük az adatokat, mint amivel a tanítási folyamat végbement, a rendszer teljesítménye romlik, néhány esetben akár drasztikusan is.

2. A FELISMERŐ TANÍTÁSA TISZTA ÉS ZAJOS BESZÉDDEL

Számos módszert megvizsgáltak már a zaj csökkentésére és a beszéd kiemelésére. Ezek a módszerek általában két alapelvet alkalmaznak. Egyik a becsült zajszint időtartománybeli kivonásán alapszik, a másik pedig a beszédjel szűrésén.

Másik módja a hatékonyság növelésnek, a zajos beszéddel való tanítás. Ezek a módszerek olyan környezetben alkalmazhatóak, ahol egy bizonyos zaj fordul elő, mint a visszhang az irodákban vagy taxiban a forgalom zaja, stb.

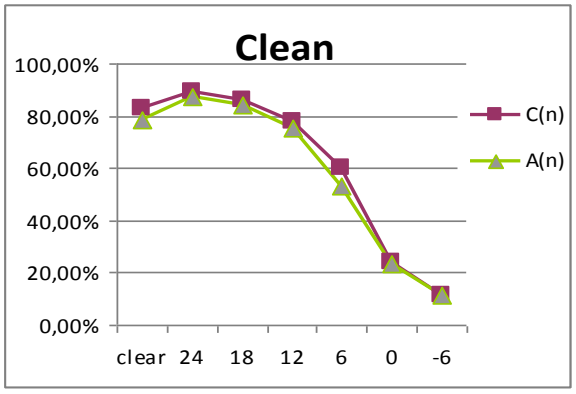
A beszédjel szűréséhez általában Wiener szűrőt vagy még nagyobb általánosságban tekintve illesztett szűrőt (matched filter) alkalmaznak. Ezek az eljárások csak alkalmasak a zajcsökkentésre, ha a zaj stacionárius. A valóságban viszont ez a feltétel a legtöbb esetben nem teljesül. A probléma ugyanaz, mint annál az általános technikánál, amit már fentebb is említettünk (ahol a felismerőt ugyanolyan zajterheléssel tanítanak be, mint ahol alkalmazzák azt)[1]. A módszer alapján azt feltételezhetjük, hogy a környezet ismert és nem változik az idő múlásával.

Az 1. ábra szemlélteti az első kísérletünk eredményeit. A vizsgálathoz egy kulcsszó alapú felismerőt alkalmaztunk.

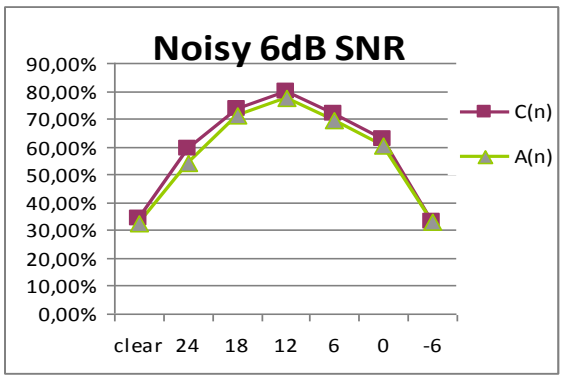
* *Automatizálási és Kommunikáció-technológiai Tanszék vezetője, Miskolci Egyetem*
** *doktorandusz, Miskolci Egyetem*

A kulcsszó alapú beszédfelismerő fontosabb jellemzői:

- 50 beszélő, 26 db szó, 21 db kifejezés.
- Beszélőfüggetlen.
- 1339 tanító szó 38 beszélőtől.
- 429 tesztelő szó 12 beszélőtől.
- 16kHz, WAV (előjeles 16 bites PCM).
- 30 állapotú HMM, 36 állapotú MFCC.



a.) Tiszta beszéddel tanított beszédfelismerő



b.) zajos beszéddel tanított beszédfelismerő, 6dB jel/zaj viszony

1. ábra HMM alapú beszédfelismerő felismerési arányai különböző jel/zaj viszonyú tesztelő anyagok esetén

A beszédfelismerőt tiszta beszéddel és 6dB-es jel/zaj viszonyú (SNR) Gauss zajjal terhelt beszéddel tanítottuk. Az utóbbi zajos beszédre hatékonyabban működött, mint a tiszta beszéddel tanított felismerő. Ugyanakkor a hatékonyság a kissé jobb minőségű beszédnél a legnagyobb. A felismerés hatékonyságát két tulajdonsággal írjuk le:

$$\%Correctness = \frac{H}{N} x100\%$$

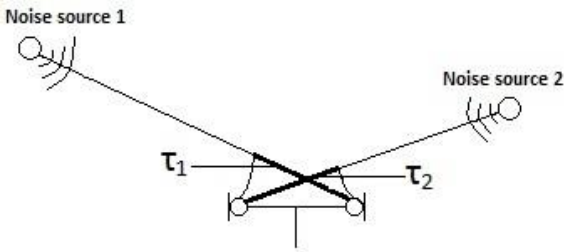
$$\%Accuracy = \frac{H-I}{N} x100\%$$

Ahol H a helyesen felismert elemek száma, I a beszúrások száma és N az összes felismerendő elem száma. $C(n)$ a zajos beszéd felismerésének a helyessége, $A(n)$ pedig annak pontossága.

A korábbi megállapításokat alátámasztják a [2] hivatkozásban olvasható vizsgálatok eredményei is. Egy kereskedelemben kapható beszédfelismerőt teszteltek (a típusát nem adták meg). A rendszert négy különböző jel-zaj szinttel tanították be: 15, 18, 21, és 24 dB, majd a 10 és 30 dB-es jel/zaj viszony közötti tartományban tesztelték. A felismerés pontossága minden esetben akkor volt a legmagasabb, ha a felismerendő beszéd jel/zaj viszonya kicsivel magasabb volt, mint a tanító jel/zaj viszonya, a többi SNR szintnél pedig romlott a pontosság. Példánkban a felismerőt 6dB-es jel/zaj viszony mellet tanítottuk, a felismerési arány maximumát pedig 12dB zajszintnél érte el, és drasztikusan lecsökkent tiszta beszéd esetén. A többi zajszint vizsgálatakor hasonló eredmények születtek.

3. ZAJCSÖKKENTÉS TÖBBMIKROFONOS MÓDSZERREL

A többmikrofonos módszer az egy mikrofonos technikák alapelveit alkalmazza némileg módosított formában, és robusztus zajcsökkentő rendszert testesít meg. Ebben az esetben a térbeli információ felhasználható az irányított minták szűréséhez, így növelve a teljes rendszer hatékonyságát. Kísérletünkben kétmikrofonos rendszert modelleztünk, amit különböző additív zajok rontanak le. A jelek sztereó fejmikrofonnal lettek rögzítve. Az első csatorna mikrofonja a beszélő felé irányul, míg a másik mikrofon sokkal kevésbé érzékeli a beszédet. A zajt mindkét mikrofon érzékeli valamilyen késleltetéssel, ami a zajforrás elhelyezkedésétől függ. A 2. ábra két eltérő zajforrás elhelyezkedését szemlélteti és a késleltetés különbsége a



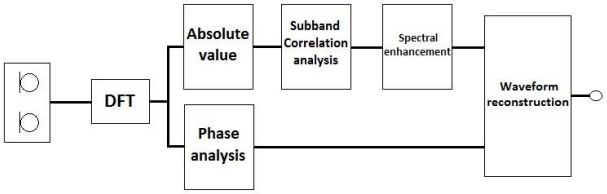
mikrofonoktól való eltérő távolságból adódik.

2. ábra Késleltetési idők szemléltetése zajforrások eltérő elhelyezkedése alapján

Amint láthatjuk, a két mikrofonba érkező zaj időbeli eltolódásának az előjele is eltérő a zajforrások elhelyezkedéséből adódóan. A helyzet pedig folyamatosan változhat, ha a mikrofonok és a zajforrások mozgásban vannak. Ez az oka annak, hogy a korábbi rendszerek - ahol megkísérelték kivonni a zajt az időtartományban - nem voltak eredményesek [3].

A τ késleltetés $e^{-j\omega\tau}$ szorzót eredményez a jel Fourier transzformáltjában, így a spektrumból való kivonás ígéretesebb megoldásnak tűnik, mivel a szorzótag nem módosítja a spektrum abszolút értékét [4, 5]. Problémát okoz, hogy ha vesszük a spektrum abszolút értékét, elveszítjük a beszéd hullámformájának visszaállításához szükséges fázis információt. Ez a tény megköveteli, hogy a Fourier transzformáció elvégzése előtt megvizsgáljuk a jel fázisát. (Néhány lényegkiemelési módszer a spektrum abszolút értékét használja, ezen metódusoknak nincs szükségük a fázis előzetes vizsgálatára.)

A másik probléma, amivel meg kell birkóznunk, a mikrofonok eltérő irány karakterisztikája, ami megmutatja, hogy mennyire érzékeny az eltérő irányból érkező hangokra. Ez a jelenség felveti a zajból fakadó különböző csillapítások és korrelációk problémáját is [6]. Zajforrások feltérképezése két mikrofonnal nagy kihívás. Esélyesebb, ha megpróbáljuk meghatározni a csillapítást a két jel részsávokra való szétválasztásával és a keresztkorreláció kiszámításával sávról sávra. A csillapítás modellezéséhez, mindkét csatorna jeléhez zajt adunk eltérő késleltetéssel és amplitúdóval.



3. ábra A zajcsökkentő rendszer felépítése

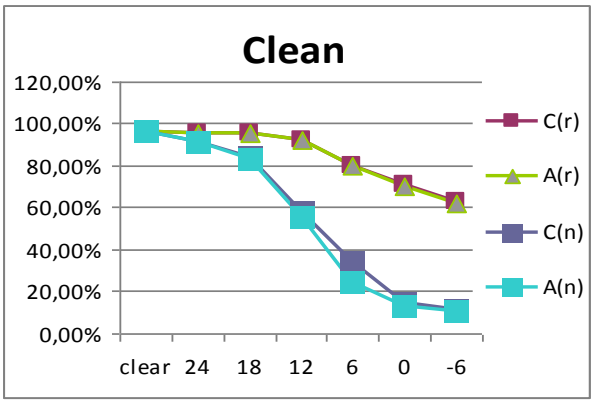
Eltérően a zajos beszédtől, ami tartalmazza a nem kívánatos zajokat, elvégezve a zajcsökkentést (zajszűrést) a beszédjel – amplitúdó és fázis – torzulást szenved.

4. KULCSSZÓ FELISMERÉS

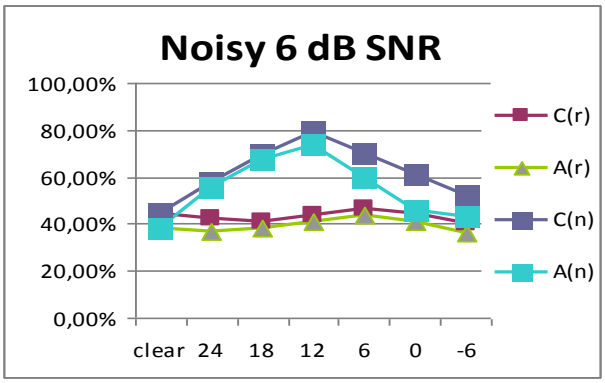
Ahogy az a 4. ábrán látható, abban az esetben, ha a rejtett Markov modell (HMM) alapú felismerőt tiszta és zajos beszéddel tanítjuk, a felismerési arányok zajszűrt beszéd felismerésénél a tiszta beszéd szintjétől kezdve kezdenek romlani. Minél jobb a tanító anyag minősége, annál nagyobb a felismerési arány tiszta beszédnél és annál meredekebb a romlás, minél magasabb a jel/zaj viszony.

A görbék jelentései:

- C(r) – zajszűrt beszéd felismerésének pontossága,
- A(r) – zajszűrt beszéd felismerésének helyessége,
- C(n) – zajos beszéd felismerésének helyessége,
- A(n) – zajos beszéd felismerésének pontossága.



a.) *Tiszta beszéddel tanított beszédfelismerő*



b.) 6dB jel-zaj szinttel terhelt beszéddel tanított beszédfelismerő

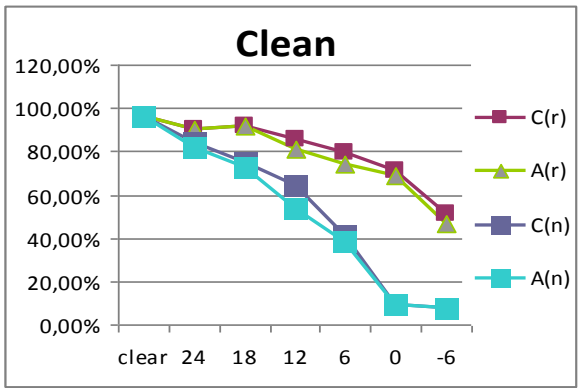
4. ábra Kulcsszó alapú beszédfelismerő felismerési arányi kalapács ütés zajjal terhelt tesztelő anyagok esetén

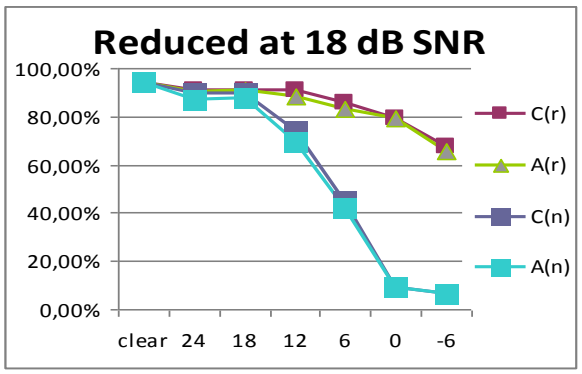
Ellentétben a zajos beszéddel tanított HMM-ek felismerési értékeivel, amikor a felismerőt zajszűrt mintákkal tanítjuk, a felismerési arányok nem romlanak a kevésbé zajos vagy tiszta beszédnél (5. ábra). Zajszűrt beszéddel tesztelve a rendszert, a felismerési arányok romlása a nagyon zajos beszédnél is jelentősen csökken.

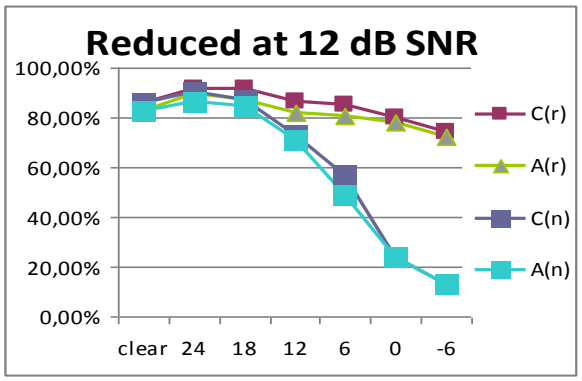
A felismerőt négy különböző zajmintával vizsgáltuk:

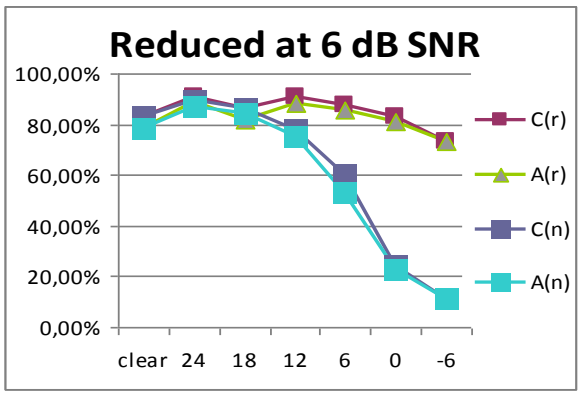
- Gauss (G),
- motoros fűrész (P),
- kalapácsütés (H), és
- lemezfeldolgozás (S) iparból származó zajokkal.

Összehasonlítottuk a felismerési értékeket stacionárius (G) zajjal terhelt majd zajszűrt beszédnél és a nem stacionárius (többi) zajjal terhelt és zajszűrt beszédnél. A felismerési arányok azt mutatták, hogy módszer alkalmazásával a teljesítmény hasonló a stacionárius Gauss zaj és nem stacionárius ipari zajok esetén. A zajmikrofonnal kiegészített rendszer tehát képes a zaj hatásának hatékony csökkentésére a nem stacionárius zajok esetében is, amit egymikrofonos rendszerrel nem tudunk megvalósítani.









5. ábra Kulcsszó alapú felismerés értékei Gauss zajjal terhelt beszédfelismerésnél, a tanító anyag zajmentes, 18dB, 12dB és 6dB zajjal terhelt (fentről lefelé)

5. ÖSSZEGZÉS

Kifejlesztettünk egy spektrális kivonáson alapuló zajcsökkentő módszert fázis rekonstrukcióval kétmikrofonos zajszűréshez. A zajcsökkentés egyformán hatékony volt stacionárius és nem stacionárius additív zajok esetén, és akkor adta a legjobb eredményt, amikor zajszűrt beszéddel volt tanítva és tesztelve. Egy automatikus beszédfelismerő kevésbé érzékeny a zajszűrés okozta torzulásokra, mint magára a zajra. Zajszűrés után a HMM alapú beszédfelismerő felismerési arányai a tanítás után olyan magasak, mint amit egy 12-18 dB-lel jobb jel/zaj viszonyú zajszűrés nélküli beszédnél mértünk.

5. KÖSZÖNETNYILVÁNÍTÁS

A bemutatott kutató munka a TÁMOP-4.2.1.B-10/2/KONV-2010-0001 jelű projekt részeként az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

6. IRODALOM

[1] DAUTRICH-B-A. RABINER-L-R. MARTIN-T-B.: On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition. IEEE Transactions in Acoustics, Speech and Signal Processing ASSP-31, pp. 793-806. 1983.

[2] FRIED-N. CUPERMAN-V.: Evaluation of Speech Recognition Equipment in a Vehicular Environment. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing. pp. 455-458. 1-2 June 1989.

[3] DAL DEGAN-N. PRATI-C.: Acoustic Noise Analysis and Speech Enhancement Techniques for Mobile Radio Applications. Signal Processing. vol. 15, pp. 43-56. 1988.

[4] DAVÍDEK, V., SOVKA, P., ŠIKA, J.: Real-time implementation of spectral subtraction algorithm for suppression of acoustic noise in speech. In Proceedings of the 4th European Conference on Speech Communication and Technology, EUROSPEECH '95, Madrid, pp. 141-144, September 1995.

[5] QUANG HUNG P., PAVEL S.: A Family of Coherence-Based Multi-Microphone Speech Enhancement Systems, Radioengineering, VOL. 12, NO. 2, June 2003

[6] RICHARD C. HENDRIKS AND TIMO G.: Noise Correlation Matrix Estimation for Multi-Microphone Speech Enhancement, IEEE Transactions on Audio, Speech, And Language Processing, VOL. 20, NO. 1, January 2012