A digitális beszédfeldolgozás újabb eredményei: beszédkódolás, beszédfelismerés és beszédszintézis

HELMUT MANGOLD
AEG Forschungsinstitut Ulm

ÖSSZEFOGLALÁS:

A beszédfeldolgozás az utóbbi években jelentős előrehaladást tett a digitális jelfeldolgozás lehetőségeinek lényeges javulásával. Ez érvényes mind a digitális beszédkódolásra és beszédátvitelre, mind az automatikus beszédfelismerés és beszédszintézis területére.

Az automatikus beszédfelismerés és beszédszintézis már közben olyan szintet ért el, hogy bizonyos területeken érdemes alkalmazni őket. Ezáltal az ember-gép párbeszéd könnyebbé és megbízhatóbbá vált.

MANGOLD, H.

A Müncheni Műszaki Egyetemen híradástechnikát és informatikát tanult. 1962-től a Telefunkennél dolgozott különféle fejlesztési munkákon. 1964-től az AEG-Telefunken ulmi kutatóintézeténeél, a beszédlaboratóriumban a beszédkódolás új eljárásainak vizsgálatával foglalkozott. 1967-ben laborvezetővé nevezték ki, és mint új témát az automatikus beszédfelismerést és beszédszintézist vezette. 1975 óta az AEG kutatóintézetében azt az osztályt vezeti, mely a beszédjelek feldolgozásával, felismerésével és szintézisével, valamint képanalízissel és kódolással foglalkozik. Ehhez csatlakozik még a digitális audiótechnika témakörének kutatása.

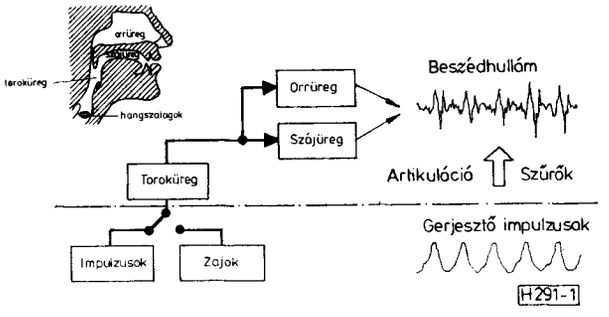
1. Beszédjel

A beszédjelek, az emberi beszédtraktusban létrejöttüktől kezdve, egészen jellegzetes jelek, melyekben az információ specifikus módon van kódolva. A feldolgozó algoritmusoknak ezeket a sajátos tulajdonságokat figyelembe kell venniük. Csak ebben az esetben lehetséges az információtartalom hatékony feldolgozása és elemzése.

A természetes beszédelőállításnál a gégében vagy a beszédtraktus szűkületében gerjesztőjelek képződnek, melyeket azután a torok, a száj- és az orrüreg az adott állásának megfelelően akusztikailag megszűr (1. ábra). Ezt a folyamatot artikulációnak nevezzük.

A beszédjel információjának lényeges elemei a jelek különböző jellemzőiben találhatók:

* Az időfüggvény periodikus szerkezete zöngés hangokat, pl. magánhangzókat jelez.
* Zajszerű jelrészletek zöngétlen hangokra utalnak.
* A jelsprektrum burkolója jellemzi az artikulációs szervek állását és ezzel a speciális hangot.



1. ábra. A beszédjel természetes előállításának elve

Fordította: Koutny Ilona
Elhangzott az 1987. máj 6—7-én tartott VDE konferencián.

* A tulajdonképpeni beszédfolyamat mindezeknek a paramétereknek a dinamikus változásában rejlik.

A digitális jelfeldolgozás sokféle nagyon hatékony feldolgozási algoritmushoz vezetett, melyek lehetővé teszik a beszédjel különböző paramétereinek elemzését.

Lényegében megkülönböztetjük a tisztán jelelemző eljárásokat és azokat, melyek osztályozással vannak egybekötve és ezért erősebben kapcsolódnak a jel tulajdonképpeni információtartalmához.

Az előbbi eljárásokhoz tartozik például a szűrés, a transzformációk különböző formái és az autokorreláció; az osztályozás-orientáltakhoz pedig a szegmentálás, az alapfrekvencia-elemzés, hangosztályozás, valamint a jelfelismerés egész területe. Példaként egy beszédjel digitálisan kiszámított spektrumát mutatjuk be a második ábrán.

2. Digitális beszédkódolás

A digitális beszédkódolás feladata, hogy a kommunikáció számára fontos információtartalmat úgy felkészítse, hogy egy rákövetkező átvitel és dekódolás ismét lehetőleg torzulatlan jelet eredményezzen, amely jól érthető és természetesen cseng. Mindamellett valójában csak releváns információt kell átvinni, a kódolónak el kell hagynia a redundáns és irreleváns információkat.

Digitális beszédkódoló rendszereket időközben nem csak híradástechnikai átvitelben használnak, hanem már a közeljövőben fontos szerepet játszanak a beszédtároló szolgáltatások, akár az emberek közötti kommunikációt megkönnyítő hangos posta (Voice-Mail), akár a számos bemondó és információszolgáltató rendszer terén.

A kódoló elv és a szükséges adatátviteli sebesség alapján lényegében három különböző kódolási

csoportot különböztetünk meg: a hullámforma-kódolást, a paraméteres kódolást és a keverteket, melyeknél mindkét elv elemeit alkalmazzák. Kiválasztott példák alapján bemutatjuk a különböző lehetőségeket.

2.1 *Hullámforma-kódolás* [1]

A hullámforma-kódolók az időfüggvényben rejlő statisztikus redundanciát használják ki.

A hullámforma-kódoló legismertebb és legegyszerűbb példája a kompressziós pulzus-kód-moduláció **(PCM)**, amelynél a minták nemlineáris kvantálása egy nagy dinamikai tartományban konstans jel/zaj viszonyhoz vezet. A kódolásnak evvel az egyszerű formájával a beszédet az időközben 64 kbit/s-ra normált átviteli sebességgel jó minőségűen lehet átvinni.

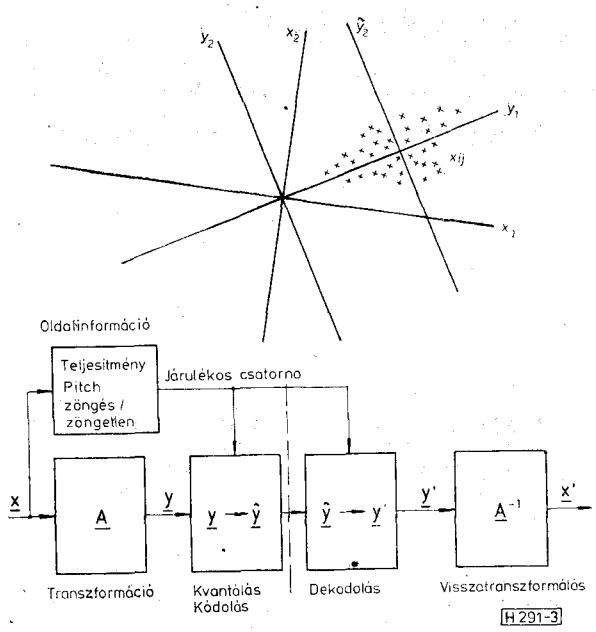
A digitális kódolásról szóló újabb munkák egyre inkább a 8 kbit/s-tól mintegy 16 kbit/s-ig terjedő középső átviteli tartománnyal foglalkoznak. A kapcsolt rádiótelefon adáshoz tervezett digitális adórendszereket ennél az átviteli sebességnél igen hatékonyan lehet üzemeltetni.

A hullámforma-kódoló modern példájaként a transzformáció-kódolót mutatjuk be. A **PCM** és számos változatával szemben, itt a képtartománybeli, azaz pl. a frekvenciatartománybeli, jelstatisztikát használják ki [2]. A 3. ábra mutatja példaként a mintavételek statisztikus eloszlását kétdimenziós koordinátarendszerben. Az eredetileg x_1-gyel és x_2-vel leírt mintákat egy elforgatott és eltolt y_1 és y_2 koordinátarendszerben jelentősen kevesebb bittel lehet leírni.

A kódolandó jel mintáit egy megfelelő jeltérben egy transzformációnak (pl. Fourier-transzformáció) vetik alá, és ott kódolják és továbbítják. A vevőhelyen történik a dekódolás és a visszatranszformálás. Általában egy kiegészítő csatornán a beszédjel további jellemzőit, mint például a zöngés hangrészletek periodicitását, értékelik ki és továbbítják. Ezzel a transzformált jel kódolása tovább javítható.

A transzformáció-kódolással minden nehézség nélkül elérhető 10—20 kbit/s átviteli sebesség.

2. ábra. A „Lautsprache" szó spektrogramja



3. ábra. Transzformáció-kódoló; fent: koordináta-transzformáció; lent: elvi megvalósítás

Ezeknél az arányoknál a transzformáció-kódoló megfontolandó az átviteli minőséget illetően más eljárásokkal szemben [3].

2.2 *Paraméteres rendszerek*

A paraméteres beszédkódolásnál a beszédjel nem bármilyen formában lesz közvetlenül kódolva, hanem egy modell szerint, amelynek paraméterei adottak. Egy ilyen modell-elképzelés általában a természetes emberi beszédelőállításra irányul. Így kapjuk a 4. ábra szerinti kódoló elvet.

Párhuzamosan mind a gerjesztőjelet, mind a beszédtraktus átviteli függvényének a karakterisztikáját közlik. Ez a két paraméter jellemzi, az 1. ábra szerint, a beszédelőállításhoz szükséges összes tulajdonságot. Evvel lehet irányítani a vevőnél végbemenő, a szintézishez szükséges folyamatokat. Ezért nevezzük az ilyen rendszereket analízis-szintézis rendszereknek is.

Így a szükséges átviteli sebesség egészen 2 kbit/s-ra csökkenthető. A vocoderek az egyedüli olyan rendszerek, amelyekkel a digitális beszédátvitel rövidhullámon is lehetséges.

A vocoderek néhány éve egy rendkívül érdekes, új alkalmazásra találtak a digitális beszédtárolás területén. Megfelelő beszédminőségnél is alacsony tárolófelhasználásuk miatt széleskörűen bevetik őket az információ-szolgáltató rendszerekben és a hangos posta rendszerekben is.

2.3 *Kevert kódolás*

Beszédjelnél alkalmazott vegyes kódolón olyan kódolót értünk, melynél mind jelstatisztikai tulajdonságok, mind pedig a beszédmodellnek vagy a hallásunk működésének tulajdonságai felhasználásra kerülnek a kódolásnál.

Erre példa a részsáv-kódoló [4]. A részsáv-kódolók viszonylag szerény hardware-ráfordítást

igényelnek, és ennek ellenére meglepően jó beszédminőséget nyújtanak.

A beszédfrekvenciasávot több részsávra osztják. Ezek a sávok a különböző mértékű hallásérzékenységnek megfelelően a nagy frekvenciáknál sokkal szélesebbek. Az egyes sávokat így különböző frekvenciákkal tapogatják le, mégpedig a magasabb sávoknál az alsó sávszélnél. A felső sávok kvantáló pontosságát erősen lehet csökkenteni. Így egy olyan összadathalmaz adódik, amely jóval kisebb, mint a PCM-nél. A beszédjel természetessége mindamellett messzemenően megmarad.

3. Automatikus beszédfelismerés

Míg néhány évvel ezelőtt a digitális beszédkódolás állt a tudományos és technikai érdeklődés előterében, ma egyértelműen az automatikus felismerésnek és szintézisnek van prioritása. Az ember és gép közötti párbeszéd egyszerűsítését nem kis mértékben a gépi beszédfeldolgozás teszi lehetővé.

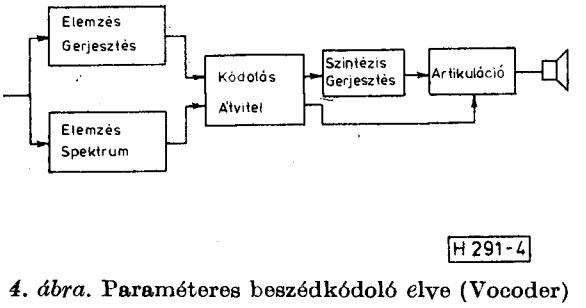
3.1 *Szófelismerő*

Napjainkban az alkalmazható beszédfelismerők gyakorlatilag mindannyian egész szavak alapján dolgoznak, azaz a legkisebb felismerési egység a szó. Ezzel gyakorlatilag már sok felmerülő feladatot, például az ipari technikában vagy irodákban, meg lehet oldani.

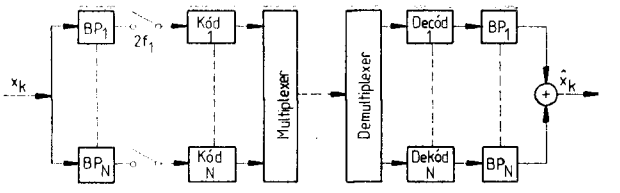
A beszédjelből paramétereket vonnak ki, rendszerint a jelspektrumot (2. ábra). A felismerendő szó kezdetének és végének a meghatározása után egy előfeldolgozási fázisban különböző normálások történnek. Ezek az előfeldolgozási lépések arra szolgálnak, hogy a különböző ejtésváltozatok sokféleségét valamennyire redukálják.

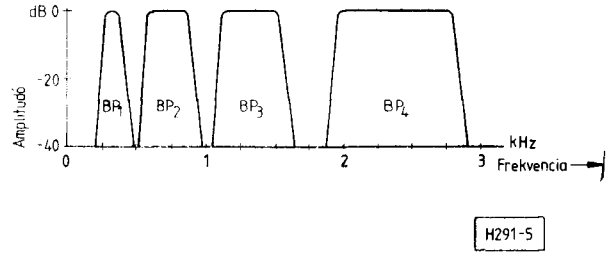
Ezután kezdődik a tulajdonképpeni felismerés, azaz a jelosztályozás. Az osztályozásra időközben a dinamikus programozás módszere honosult meg. Ennél az osztályozandó mintát az idő-tengely mentén leképezik a referenciamintára úgy, hogy optimális legyen a megfeleltetés. Az alkalmazott időtorzítás mértéke, valamint a két minta, ezután mért spektrális euklideszi távolsága szolgáltatja a mértéket a két minta hasonlóságára. Egy utólagos szintaktikai feldolgozás javíthatja a felismerési eredményt.

Az ilyen izoláltszavas felismerők kielégítő megbízhatósággal leginkább csak beszélőfüggően működnek. Tehát minden felismerendő szót a rendszernek előzőleg be kell mondani. Így a referenciaminták automatikusan létrejönnek, de természetesen beszélőspecifikusak. Kisebb szókincs, mintegy 200—500 szó esetén elérhető a 99% feletti helyes felismerési arány [5].



4. ábra. Paraméteres beszédkódoló elve (Vocoder)





5. ábra. A részsáv-kódolás elve

Egészen más a helyzet beszélőfüggetlen rendszereknél. Itt az eddig lehetséges felismerési arányok jóval alacsonyabbak. Laboratóriumban mért felismerési arányok 95% körül vannak.

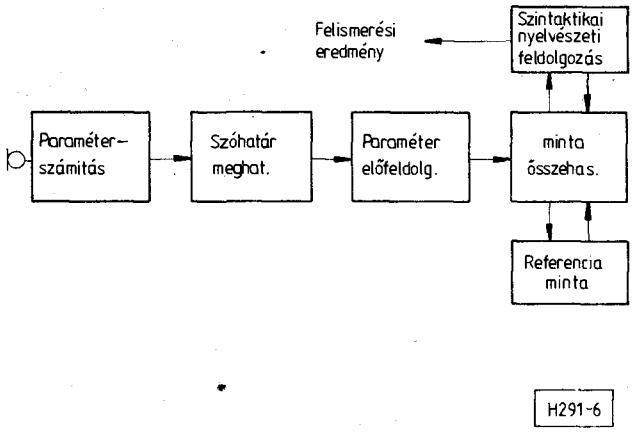
Hasonló problémák adódnak magas zajszintnél. Jóllehet izoláltszavas felismerőket gyakorlatilag már sok feladatra alkalmaznak, mégis intenzív alapkutatásra van szükség az elkövetkezendő években.

3.2 *Kapcsolt szavak felismerése*

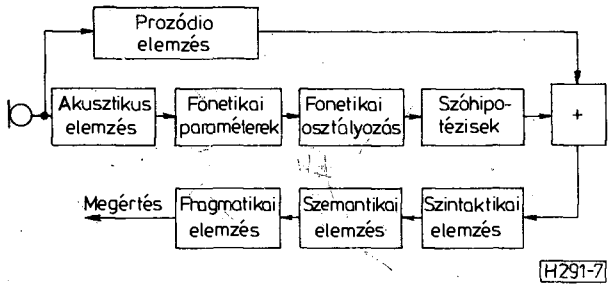
Egy ideje intenzív fejlesztéseket végeznek kapcsoltan kimondott szavak megbízható felismerése terén. Kapcsoltszavas beszédfelismerő alatt még nem folyamatos beszédfelismerőt értünk, azaz nem a szokásos mindennapi beszédünk felismerését. Kapcsoltszavas felismerők messzemenően az izoláltszavas felismerő elvén működnek. Tehát elvüket tekintve ugyanúgy épülnek fel, mint az izoláltszavas felismerők (6. ábra).

Megkülönböztetünk explicite és implicite szegmentáló felismerőket [6], ahol az osztályozás és szegmentálás ugyanabban a feldolgozási fázisban megy végbe. Mindkét eljárásnak a teljesítőképessége körülbelül ugyanolyan.

Számcsoportokkal — akár 5 egymás után következő számmal — végzett tesztek kb. 97%-os csoportfelismerési arányt produkálnak beszélőadaptív üzemben. Itt még jobban megmutatkozik, mint az izoláltszavas felismerőknél, hogy egy teljesítőképes tanulóeljárás a felismerés minőségére egészen döntő. A megtanulandó referenciamintába bele kell dolgozni egy szónak a különböző ejtéseit, a szócsoport elején, közepén és végén.



6. ábra. Izoláltszavas felismerő alapvető felépítése



7. *ábra.* Beszédmegértő rendszer elvi vázlata

3.3 *Beszédmegértés*

A gépi beszédfelismerés területén minden kutatási munka tulajdonképpeni célja természetesen olyan rendszerek kifejlesztése, melyek teljesítőképessége egyre inkább megközelíti az emberét a beszéd megértésében. Hosszú az út odáig, és nem is teljesen világos, hogy ezt a célt egyáltalán el lehet érni. Mégis egy egész sor olyan köztes célkitűzés van, melyeket addig kutatási projektekben vizsgálni lehet [7].

Beszédmegértő rendszerek már nem dolgozhatnak egész szavak felismerése alapján. Egyedül az a tény, hogy folyamatos beszédben sok szót koartikulálunk, ezeket nem egyenként, hanem folyamatosan ejtjük ki, lehetetlenné teszi a szó, mint egység alkalmazását. Ehelyett hangokat vagy hangkapcsolatokat alkalmaznak a felismerés alapelemeként.

Egy beszédmegértő rendszerben több, hierarchikus lépcsőben hajtják végre az egyes elemző lépéseket. A 7. ábra durva áttekintést ad a folyamatról.

A jelosztályozás kiértékelt szóhipotézisek felállításával fejeződik be. Ezek a hozzákapcsolt magasabb feldolgozási lépcsőkben tovább elemződnek fonetikai és szemantikai összefüggésben.

4. Automatikus beszédszintézis

A gépi beszédfelismeréssel ellentétben, mely a többévi intenzív kutatás ellenére is csak szerény eredményeket tud felmutatni, a beszédszintézist különböző formáiban intenzíven alkalmazzák. A siker részben annak is köszönhető, hogy a szintetizált jel vevője mindig az ember, aki egy nem teljesen természetes jelet is meg tud érteni az emberi percepciós rendszer magas teljesítménye következtében.

4.1 *Beprodukáló eljárás*

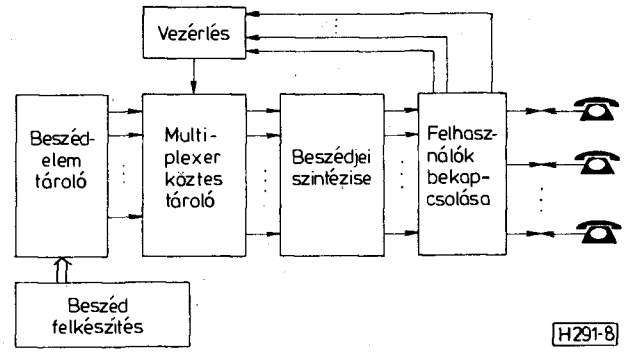
Reprodukáló vagy félig-szintetizáló eljárás alatt olyan beszédkimenetű rendszert értünk, ahol a kiadandó szöveg több vagy kevesebb olyan elemből tevődik össze, melyeket rendszerint előzőleg egy emberi beszélő bemondott és azután digitálisan tároltak.

A félig-szintetizáló rendszerek nem mások, mint rugalmasan irányítható tárolórendszerek. A minőség végső soron attól függ, hogy a báziselemeket milyen ügyesen választották ki és készítették elő, és hogy a jelkódolás melyik formáját használják. A technika mai állásánál már nem probléma a telefonnál megszokott minőséget ezeknél a rendszereknél is biztosítani.

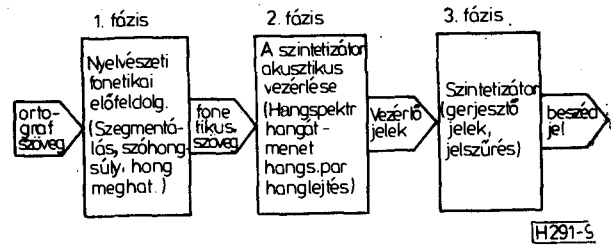
Félig-szintetizáló rendszereket akkor lehet előnyösen alkalmazni, ha a kiadandó szöveg előzőleg ismert, és csak nagyon ritkán fordulnak elő szövegváltozások, mivel minden szövegváltozás nagy mennyiségű manuális előkészítő munkát igényel. Amennyiben gyakorlatilag bármilyen szöveget, amely gyakran változhat, akarunk kiadni, akkor a teljes szintézis a megfelelő rendszer.

4.2 *Beszédelőállítás teljes szintézissel*

A teljes szintézissel dolgozó rendszerek a beszédjel előállításához nem igényelnek tárolt beszédelemeket, hanem az ortografikusan írt szövegből teljes egészében beszédet tudnak generálni. Ezért ango-



8. ábra. Félig-szintetizáló beszédelőállító rendszer elve



9. ábra. A teljes szintézis alapelve

lul *text-to-speech* rendszereknek nevezik őket. Lényegében három feldolgozási lépést tudunk megkülönböztetni.

A nyelvészeti fázisban lesznek a beadott szövegből kivonva és elemezve a kiejtés számára fontos jellemzők egy átfogó és nyelvspecifikus szabályrendszer segítségével.

A következő fázis tárolja az artikulációs szabályokat és meghatározza a tulajdonképpeni beszédszintetizátor vezérléséhez szükséges vezérlő paramétereket. Itt hozzák létre például a hangátmeneteket, melyeknek lehetőleg természetesen kell csengeniük.

A harmadik fázis állítja végül elő a tényleges beszédjelet a gerjesztő és átviteli függvényből.

Az első jól érthető teljes-szintézisű rendszerek nem régóta vannak a piacon. Már különböző berendezésekben alkalmazzák őket. A teljes szintézist először a testi fogyatékosokat segítő segédeszközökben használták fel, elsősorban vakoknak szánt felolvasógépekben. Időközben a teljes szintézist irodai rendszerekben is egyre inkább alkalmazzák, hála a megjavult beszédminőségnek.

5. Kitekintés

A digitális beszédfeldolgozás az elmúlt években nagymértékben szélesítette a kommunikáció lehetőségeit az emberek számára. Mind az emberek közötti, mind az ember-gép kommunikáció terén olyan alkalmazások adódnak, melyek a pusztán híradástechnikai alkalmazáson túl az információtechnika fontos közegévé teszik az emberi beszédet. A jövőbeli kutatások segíteni fognak, hogy a beszédkommunikációban rejlő információs folyamatot jobban megértsük és ezáltal jobban felhasználhassuk.

IRODALOM

[1] *R. E. Crochiere, J. M. Tribolet:* Frequency Domain Techniques for Speech Coding, J. Acoust. Soc. Am. Dec. 1979 512—530.

[2] *R. Zelinski:* Ein System zur adaptiven Transformationscodierung mit cepstraler Steuerung und Entropiecodierung, Frequenz 36 (1982) 7/8, 193—198.

[3] *J. M. Tribolet, et. al.:* A Comparison of the Performance of Four Low Bit Rate Speech Waveform Coders, Bell Syst. Tech. J., vol. 58, March 1979, 699—712.

[4] *U. Schneider:* Digitale Sprachübertragung mit 9.6 kbit/s über Funkkanäle, NTG-Fachberiehte Bd. 94 Sprachkommunikation, 1986, 168—173.

[5] *F. Class, R. Zelinski:* Ein Algorithmus zur Beschleunigung der dynamischen Zeitnormierung für die automatische Spracherkennung, Fortschritte der Akustik — DAGA 1984, Bad Honnef, DPG-GmbH 1984, 853ff.

[6] *F. Class, H. Mangold, R. Zelinski:* Zur Segmentierung bei der automatischen Erkennung von Wortgruppen, Frequenz 34 (1980) 5, 142—148.

[7] *H. Ney,* Automatic Speech Recognition, VDE-Tage 1987, Budapest

[8] *H. Mangold,* Prinzipien und Möglichkeiten der elektronischen Sprachausgabe, ED 85, Vol. 1, Net work 1985, 1. 3. 1—17.