Beszédadatbázisok a gépi beszédfelismerés segítésére

Vicsi Klára

1 Adatbázisok jelentősége

A gépi beszéd és beszélő felismerési eljárások lényegében két jól elkülöníthető elméleti alapra épülnek. Az egyik a szabálybázisú megközelítés (kognitív módszer), a másik a statisztikai elméleti alapú feldolgozás (információelméleti megközelítés).
--Szabálybázis alapon működnek pl. a különböző szakértői rendszerek.
--Statisztikai alapú feldolgozást használnak a „Rejtett Markov Model" („Hidden Markov Model: HMM"), vagy Neurális hálózatok (Neural Network NN) használatával megvalósuló felismerők.

A mai beszédfeldolgozási tudásszinten a gyakorlatban megvalósuló sikeres beszélő és beszédfelismerő rendszerek statisztikai alapokon működnek.

A beszéd természetére jellemző a fizikai paraméterek nagymértékű variáltsága beszélők között, egy beszélőn belül, továbbá az akusztikai környezet függvényében is. A beszéd statisztikai modelljének paraméterfelülete a beszéd nagyfokú variálhatóságát kell, hogy tükrözze, így sok dimenziójú kell legyen.
Egy pontos paraméterbecslési lépés végrehajtásához (betanítási lépés) nagyszámú minta alapján történő betanítás szükséges. E minták gyűjteményei - a szükséges jegyzetekkel, címkézésekkel és átírásokkal ellátva képezik az adatbázist.

Az adatbázisoknak tartalmazni kell azokat a megfigyeléseket, amelyek a paraméterbecsléshez szükségesek, tehát mindazokat a mintákat, amelyek egységesen lefedik a beszéd (és a környezeti zajok) variáltságát. Pl. ha egy beszédhang nincs benne a betanításhoz használt adatbázisban, akkor azt a beszédhangot soha nem fogja a gép felismerni.

A mai felismerőket csak egy meghatározott szűk felhasználási területre tudják tervezni. Pl. beszédfelismerő, amely csak egy adott nyelven, telefonon keresztül bemondott számok, szavak felismerésére alkalmas, nem ismer fel mondatokat. Az un. diktáló rendszerek folyamatos beszédet képesek felismerni megadott nyelven, jól meghatározott témakörön belül, de kizárólag csak a felhasználó hangjára működnek elfogadható pontossággal. Ezek a felismerők csak csendes környezetben működnek jól. Azok a felismerők, amelyeket zajtalan környezetbe terveztek, nem működnek zajos körülmények között. Az utcazajban működő felismerő rosszul működik, ha személygépkocsiban kívánják használni. Vagyis csak azokat a mintákat képes felismerni, amelyeket előzőleg már megtapasztalt, vagyis amelyre előzőleg be lett tanítva.
Betanítás viszont adatbázisok segítségével hajtandó végre. Ezért nőtt meg az utóbbi években a jelentőségük. Óriási pénzeket költ ma a világ adatbázisokra. A soknyelvű Európa igen nagy feladat előtt áll, hiszen minden nemzet a saját nyelvén akar bekapcsolódni a nemzetközi kommunikációba, tehát nyelvenként kell sokfajta feladatra alkalmas adatbázisokat létrehozni.
Nem biztos, hogy a mai statisztikai megközelítések nyújtják a legmegfelelőbb megoldást a gépi beszédfelismerésre, de hogy igen költségesek, az bizonyos.

2 Megjegyzések a statisztikai feldolgozáshoz

A statisztikai felismerő rendszerben egy véletlen folyamat (X) előállít egy diszkrét idejű véletlen jelet ($X_{(n)}$), ahol n a diszkrét időindex. Meg kell becsülni azokat a paramétereket, amelyek a modellt jellemzik. A becslés pontossága arányos a rendelkezésre álló megvalósulások számával. Ha minden lehetséges realizáció rendelkezésünkre állna, akkor ismernénk az X véletlen folyamat együttes felületét. Ebben az esetben a modell teljesen pontos valódi paramétereit becsüljük meg.

Megjegyzendő azonban, hogy nincs értelme ugyanazon folyamat realizációit hosszú ideig gyűjteni. Lényegében haszontalan ugyanazon beszélő hangjának többszöri rögzítése, vagy több felvétel készítése, mint amennyi az együttes felület lefedéséhez szükséges.
Pl. kísérleti tény ma már, hogy 100 megfelelően kiválasztott beszélő elégséges és hatékony a beszélőfüggés betanításhoz. Tovább növelni számukat haszontalan, sőt néha káros!!

A gyakorlatban az adatbázisok tervezésénél figyelembe kell venni, hogy az adatbázis létrehozása nem más, mint a véletlenszerű folyamat egyes megvalósulásainak (realizációinak) összegyűjtése. Bármelyik megvalósulás statisztikai tulajdonságai egybe esnek a folyamat sokfajta megvalósulásának együttes tulajdonságaival, egy adott időpillanatban (n). Vegyük a beszédfolyamat együttes felületét {x(n, l)}, ahol l jelenti a járulékos függéseket speciális dimenziókon, mint pl. beszédnél a kiejtés beszélőtől való függését. Itt n az időfüggést, l index pedig a beszédfolyamat járulékos függéseit jelenti ezektől a speciális dimenzióktól. Ilyen speciális dimenzióra vonatkozó index például az amelyiket a beszélők kiejtésének azonosítására használunk. Az X beszédfolyamat együttes felületét befedő adatbázis létrehozása megkívánja ezeknek a különböző megvalósulásoknak az összegyűjtését (x(n, l)), minden l járulékos függés esetében. Ez szükséges a megfelelően pontos becslés kialakításához.

A fentiek alapján látható, hogy a paraméter becslés pontossága, tehát a felismerés jósága lényegében a betanításhoz használt adatbázis jóságán múlik. Vagyis azon, hogy az adatbázis elemei helyesen legyenek kiválasztva, egy-egy elemből megfelelő darabszámú reprezentáns legyen, az elemek minősége megfeleljen az előírásoknak stb.

Mind a HMM mind az NN alapú felismerés stacionárius folyamatok sorozataként tekinti a beszédet. Viszont a beszédképző szervünk folyamatosan állítja elő a beszédet, a statisztikai tulajdonságok időfüggők. A beszéd modellezhető, mint részeiben stacionárius folyamatok láncolata és HMM nagyon jól alkalmazható ilyen folyamatokra. Ez viszont megkívánja az adatbázis szegmentálását olyan hullámforma részekre, amelyeknek hasonlók a statisztikai tulajdonságaik. A HMM és NN felismerési algoritmusok az adatbázis pontos fonetikai átírását igénylik. A HMM felismerők a betanításhoz igénylik a beszédfolyam szegmentálását, és a már működő HMM felismerők szegmentálásra is használhatók.

3 A beszéd variáltsága

A beszéd paraméterek számos hatás következtében megváltoznak, variáltságuknak számos forrása van, amelyek a felismerést megnehezítik. Ezek a források lehetnek pl. a hangképző szervek biológiai tulajdonságai, például személyenként változnak a hangképző szervek méretei, így azt általuk létrehozott fizikai produktum is változik.

A hangképzés folyamatosan változó mozgások összessége, amelyet a felismerésnél kvantálva használunk. A folyamatos hangképzőszervi mozgások miatt egyik hang fizikai tulajdonságai

befolyásolják az azt megelőző és követő hangok fizikai tulajdonságait. Ezt nevezik koartikulációs hatásnak.

A fizikai paraméterek variáltságát okozhatják a környezeti, akusztikai körülmények. Ilyenek pl. a zajos, zajtalan környezet, visszhangok, termek, telefonbeszéd stb.
A variáltságot okozó tényezők számos módon csoportosíthatók, mégis talán a beszélőkön belüli, és a beszélők közötti variáltság szerinti csoportosítás illik a legjobban a felismerők működési tulajdonságaihoz.

3.1 Beszélőn belüli variálhatóság

Beszélőn belüli variáltság okai: a coarticuláció, a ritmus, hangerő, hangmagasság, hanglejtés, nyomatékbeli különbségek. Megfázás igen nagymértékben megváltoztatja a hangok akusztikai paramétereit. Környezeti hatások, izgalom, meglepetés stb. szintén erősen befolyásolják a létrehozott beszéd akusztikai tulajdonságait.

1. Táblázat: Beszédfelismerő tervezésének lényeges szempontjai a beszéd paraméterek variáltságát okozó tényezők csoportosításával

környezet	zaj típusa jel/zaj viszony használati körülmények
átalakító	mikrofon telefon
csatorna	sávamplitudó torzítás visszhang

beszélők	beszélőfüggőség/függetlenség beszélő neme kor fizikai és pszichikai állapot
beszédstílus	hangszín: meleg, normál, kiabálás beszédegység, izolált szavak, folyamatos beszéd, spontán beszéd beszédsebesség: lassú, normál, gyors
szótár	specifikus vagy általános az elérhető betanító anyag jellemzése

3.2 Beszélők közötti variáltság

Biológiai tényezők pl. a beszédképző szervek méretkülönbsége, ami az akusztikai paraméterek jelentős variáltságát okozza női, férfi, gyermekhangok esetében, de egy egy csoporton belül is.
Környezeti hatások két csoportban: a statikus (teremakusztikai hatások, utózengési idő, rögzítő berendezések, stb.) és dinamikus (zaj, mikrofon pozíció stb.) hatások szintén erősen befolyásolják a beszéd akusztikai paramétereit.
Nyelvi különbözőségek szintén forrásai a beszéd variáltságának.

3.3 Beszédfelismerők, betanításához szükséges adatbázisok osztályozása a beszéd variáltsága függvényében

Beszéd variáltságot okozó tényezők hatással vannak a felismerőt megvalósítandó módszerre. Tervezéskor eszerint tudni kell, hogy milyen típusú felismerőt kell létrehozni, és a variáltságot okozó tényezőket rögzíteni kell (1. Táblázat) a következők szerint:

beszélőfüggőség: függő, független

beszélőadaptáció

beszédegység: szó, folyamatos felismerés, kapcsolt szavak

beszédtempó: lassú, normál, gyors

extra, nem nyelvi kapcsolatú hangok: nyelés, köhögés

szótárméret: felismerendő elemek száma

4 Adatbázisok

Az adatbázisok számítógép segítségével létrehozott,- tárolt és a szükséges magyarázó jegyzetekkel, címkézésekkel és átírásokkal ellátott beszédfelvételek gyűjteményei.
Rádióból, TV-ből felvett beszéd nem adatbázis. Az adatbázis lényeges tartozéka a precízen leírt dokumentáció a rögzítés technikájáról, a beszélők számáról és típusáról, a nyelvi tartalomról, oly módon, hogy az adatbázist felhasználók egyszerűen megkapják a gyűjteményre vonatkozó szükséges információt.

Nagyon sokfajta adatbázis létezik. A beszédtechnológiával foglalkozó szakemberek számára igen fontos ezeknek az adatbázisoknak az ismerete, azért hogy közülük egy meghatározott feladatra a legmegfelelőbbet tudják kiválasztani, vagy ha nincs megfelelő adatbázis, hogyan kell az adott feladathoz az optimálisat létrehozni.
Az adatbázisok különböznek egymástól abban, hogy milyen felhasználási területre készültek, mekkora a bennük gyűjtött beszéd mérete, a bemondók száma, stb.

4.1 Adatbázisok felosztása

A jelenleg elérhető adatbázisok 3 alap kategóriába sorolhatók felhasználás szerint (2. Táblázat):

Analitikus – diagnosztikus adatbázis: nyelvi és fonetikai kutatások segítését szolgálja. Ilyen pl. a BABEL (EURM0, EUROM1 adatbázis).

Általános adatbázis: nem specifikus, általános szótárakat tartalmaz, sokfajta felhasználásra alkalmas, mint például a SPECO (gyermek beszédadatbázis .

Specifikus adatbázis: olyan beszédgyűjtemény, amely meghatározott felhasználási területen készül. Különböző felismerők betanítására alkalmas adatbázis. Ilyen például a SPEECHDAT adatbázis.

Adatbázisokra jellemző, hogy milyen nyelvi egységekből épülnek fel, pl. izolált szavakból, mondatokból stb., továbbá a bemondás módja szerint lehet olvasott szöveg, spontán beszéd.

2. Táblázat: Adatbázisok felhasználási terület szerint felosztása

analitikus-diagnosztikus	alap nyelvi és fonetikai kutatások segítését szolgálja
általános adatbázis	nem specifikus, általános szótárakat tartalmaz; sokfajta felhasználási területre alkalmas
specifikus adatbázis	olyan beszédgyűjtemény, amely meghatározott felhasználási területre készül

4.2 Az adatbázisok tervezése

Az adatbázisok, vagy az adott feladathoz legjobban illeszkedő adatbázis kiválasztásánál az alábbi szempontokat kell figyelembe venni: a felvételek és a rögzítés pontos fizikai leírását, a felvett anyag nyelvi jellemzőit, az adatbázis méretét, a beszélők szocio-, lingvisztikai adatait, az adatbázis feldolgozási módját (3. Táblázat).

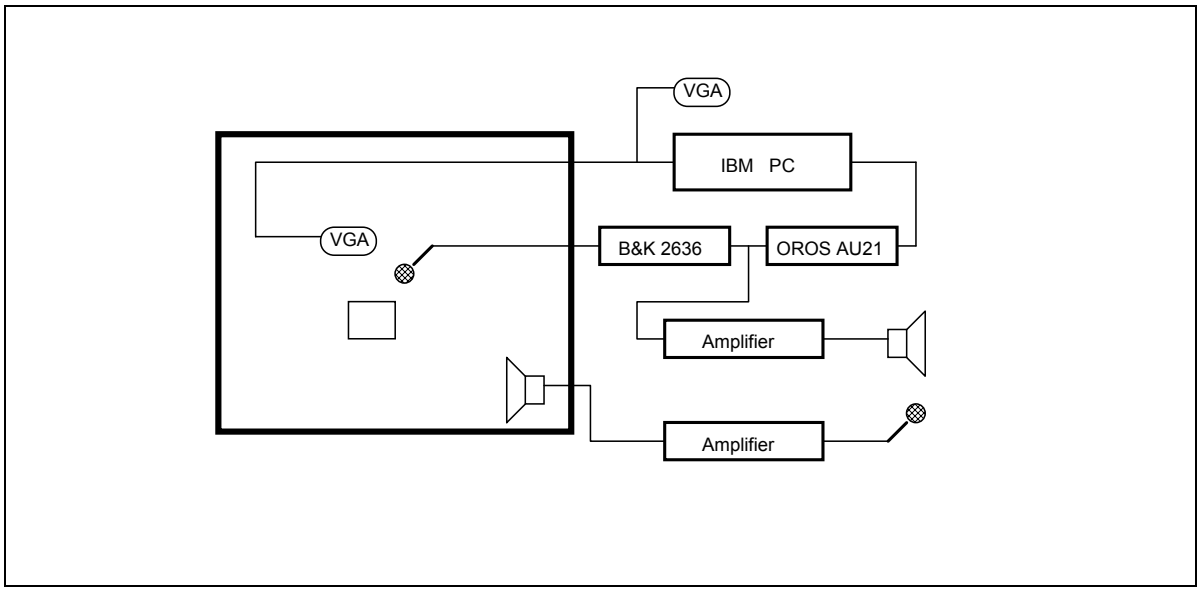
4.2.1 A beszéd adatbázis bemondási körülményei, a rögzítés módja

A felvételi körülmények pontos leírása lényeges része az adatbázisnak. Itt kell figyelembe venni:

- egy, vagy több mikrofon, mikrofon műszaki leírása
- környezet: stúdió, süketszoba, iroda stb.
- felvétel ellenőrzési módszer
- mintavételi paraméterek

Példaképpen bemutatjuk a BABEL magyar beszédadatbázis felvételi körülményeit a 2. ábrán.

2. ábra BABEL süketszobai felvételek mérési összeállítása



3. Táblázat: Adatbázisok jellemző adatai

rögzítés fizikai leírása	Mintavételi paraméterek Felvételi körülmények fizikai leírása Monitor használata
nyelvi jellemzők	Rögzített nyelv, dialektus Nyelvi alapegység: hangkapcsolatok, szavak, mondatok Bemondott anyag leírása Bemondás stílusa: olvasott, spontán beszéd, dialógus
méretbeli jellemzők	Beszélők száma Rögzített anyag időbeli hossza Nagysága CD-k száma
szocio-lingvisztikai jellemzők	nem, kor, beszéd stílusa
adatbázis feldolgozása	Címkézés Átírás Szegmentálás Spektrális elemzés

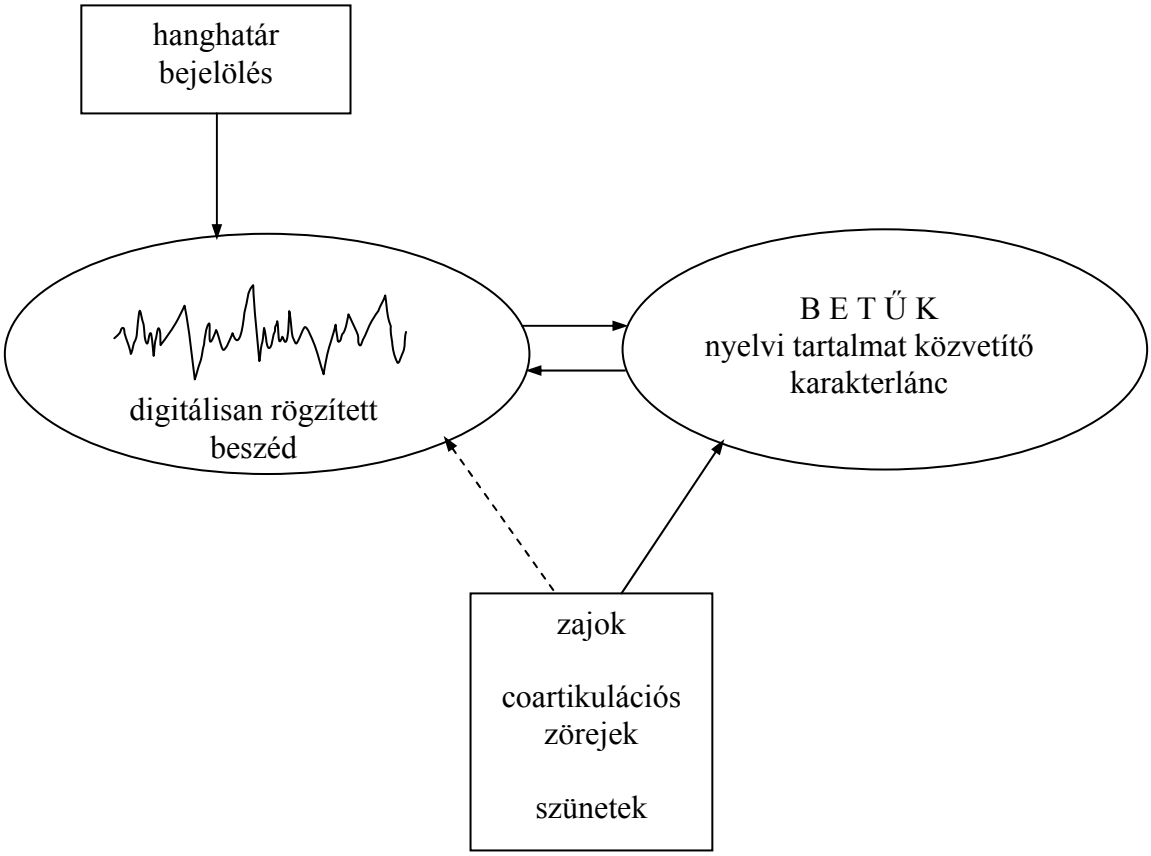
4.2.2 Méretbeli jellemzők

Beszélők száma szerint külön adatbáziscsoportok léteznek.

Kevés beszélő adatbázisa pl. beszéd szintézis fejlesztés céljait szolgálja. Lényeges jellemzője a lehető legnagyobb fonetikai variáltságú anyag összegyűjtése. Az anyagban hangsúlyozottan szerepet kapnak a beszéd mikroszegmentális jellemzői. Rendszerint a bemondást szakértő végzi.

Adatbázis közepes számú beszélővel a felismerésénél használt modell paraméterek becslésére szolgál. Éppen ezért a nyelvi szöveg variáltsága nagy. Általában csendes helyiségekben történik a felvétel. Beszélők száma kisebb, mint 50.

Adatbázis sok beszélővel: Ezek az adatbázisok a beszélő független felismerők betanítására szolgálnak. A beszédstílus, és a rögzítési körülmények nagy variáltsága szükséges.



3. ábra: Adatbázisok feldolgozása

4.2.3 Szocio-lingvisztikai jellemzők

Ebbe a csoportba tartoznak azok a jellemzők, amelyek főleg a bemondók leírására szolgálnak. Férfiak, nők, dohányoznak, nem dohányoznak. Anyanyelvükön történik-e a bemondás. Táj-szólások vannak e rögzítve az adatbázisban. Milyen a koreloszlás a bemondók között.

4.2.4 Adatbázis nyelvi feldolgozása

Beszédadatbázisoknak a beszéd digitális tárolása mellett annak nyelvi információ tartalmát is rögzíteniük kell. Ezért a hullámforma tárolása mellett a hozzátartozó ortografikus karaktereket is rögzítik.

Különböző zajok, embertől származóak,- (ilyen a köhögés, nyelés, különböző szájmozgásból adódó zajok)), vagy környezetiek (járművek, motorok zaja, székcsikorgás stb.) bejelölésre kerülnek a legtöbb adatbázisban, vagy a szöveganyagban, vagy magában az időfüggvényben (3. ábra).

4.2.4.1 Akusztikai jelek fonetikai átírása

Rendszerint a karaktereket hozzárendelik a rögzített hullámformához, vagyis a folyamatos beszédet pl. beszédhang egységekben kvantálják, bejelölik a beszédhangok elejét és végét, valamint beírják a beszédjelhez tartozó írásos szimbólumokat. Ezek a szimbólumok lehetnek egy adott nyelv betűi, de ha az adatbázis nemzetközi célra készül, akkor célszerű nemzetközi

jelölésrendszert használni, hogy a külföldi szakemberek is pontosan tudják milyen hangok sorozatáról van szó. Ilyen nemzetközi jelölésrendszer az IPhA (International Phonetic Alphabet) a múlt században született, és szimbólum rendszere több száz nyelv hangjainak leírására alkalmas. A szimbólumok még a kézíráshoz alkalmazkodtak. A mai számítógépes billentyűzetek ilyen karakterek egyszerű bevitelére nem megfelelőek, ezért a mai használat céljára a billentyűzet karaktereiből összeállított un. SAMPA szimbólumrendszert dolgoztak ki a 90-es évek elején. Az európai adatbázisok már SAMPA karaktereket használnak. A magyar beszédhangok IPhA és SAMPA szimbólumkészletét a 4. Táblázat mutatja be.

4. Táblázat: IPhA és SAMPA magyar

Betűk	Példák	IPA	SAMPA
a	hat	ɔ	O
á	hát	a:	A:
e	vet	ɛ	E
é	vét	e:	e:
i	hit	i	i
í	szít	i:	i:
o	sok	o	o
ó	sók	o:	o:
ö	köt	ø	2
ő	sőt	ø:	2:
u	fut	u	u
ú	kút	u:	u:
ü	süt	y	y
ű	fűt	y:	y:

Betűk	Példák	IPA	SAMPA
p	pad	p	p
b	bab	b	b
t	tél	t	t
d	dél	d	d
k	kép	k	k
g	gép	g	g
c	cél	ts	ts
dz	bodza	dz	dz
cs	csô	tʃ	tS
dzs	dzsem	dʒ	dZ
ty	tyúk	c	t'
gy	gyár	ɟ	d'
f	fél	f	f
v	vér	v	v
sz	szép	s	s
z	zaj	z	z
s	só	ʃ	S
zs	zsír	ʒ	Z
h	hét	h	h
r	rész	r	r
l	lép	l	l
j	jön, lyuk	j	j
m	méz	m	m
n	néz	n	n
ny	nyom	ɲ	J

Jelentősebb fonémavariációk:
/h/fonémára:

h	doh	x	x
h ch	ihlet	x̆	x
h	lehet	ɦ	h\

Zöngétlen /j/

j	kapj	ç	x˙

/m n/ fonémákra:

m	kámfor, hamvas honvágy, honfitárs	ŋ	F
n	ing tönk	ŋ	N

A fonetikai átírásnak számos szintje létezik:

Kanonikus fonetikai átírás: Az adott szöveg karaktereinek olyan átírása, amelyben az ortografikus karaktereket fonémák sorozatára alakítjuk ki, de az adott szövegkörnyezetet nem vesszük figyelembe. Tehát a hasonulás és a koartikuláció nincs figyelembe véve.

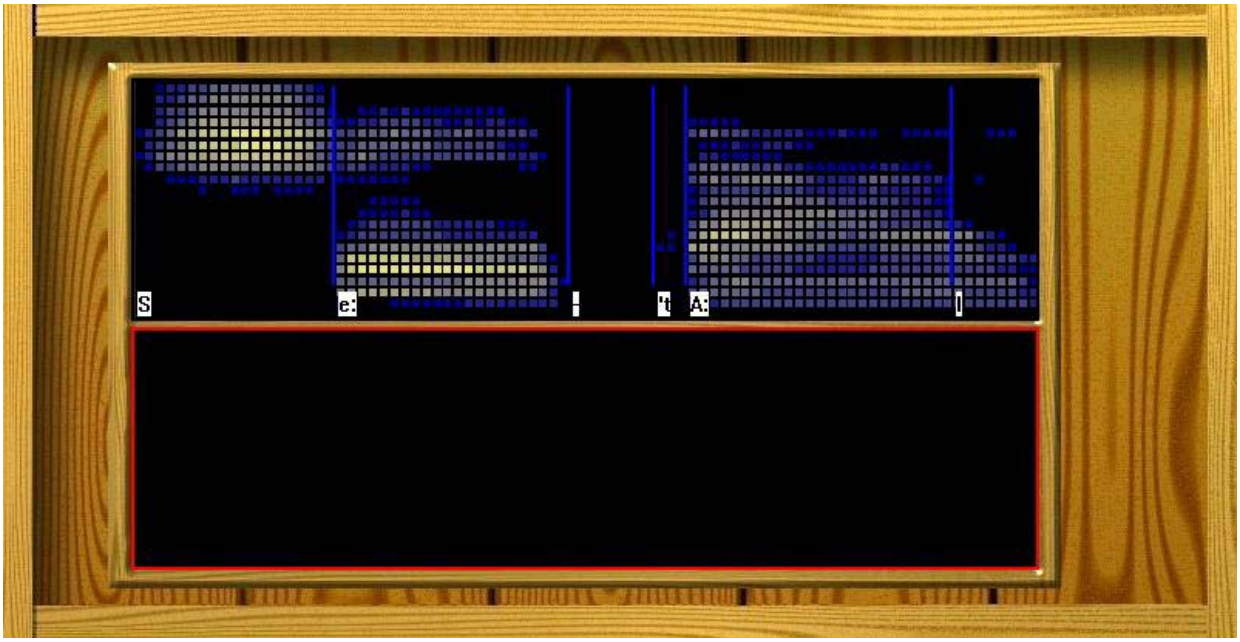
Fonotipikus fonetikai átírás: A karakterek átírását, az adott nyelv fonetikai szabályainak alapján végezzük, a szövegkörnyezet függvényében (pl. A hasonulási szabályok figyelembe vételével).

Hallás alapján történő fonetikai átírás: A figyelmesen lehallgatott szöveg hallás alapján történő lejegyzése. Tehát itt, az írott szöveg figyelembe vétele nélkül, kizárólag a hallott hangok kerülnek lejegyzésre.

Audio-vizuális fonetikai átírás: A fonémáknál kisebb egységek alapján történik az átírás, a közel stabil akusztikai-fonetikai részek bejelölésével. Az átírást a szöveg hallgatása, és az időfüggvény vagy a színkép elemzése alapján hajtják végre.

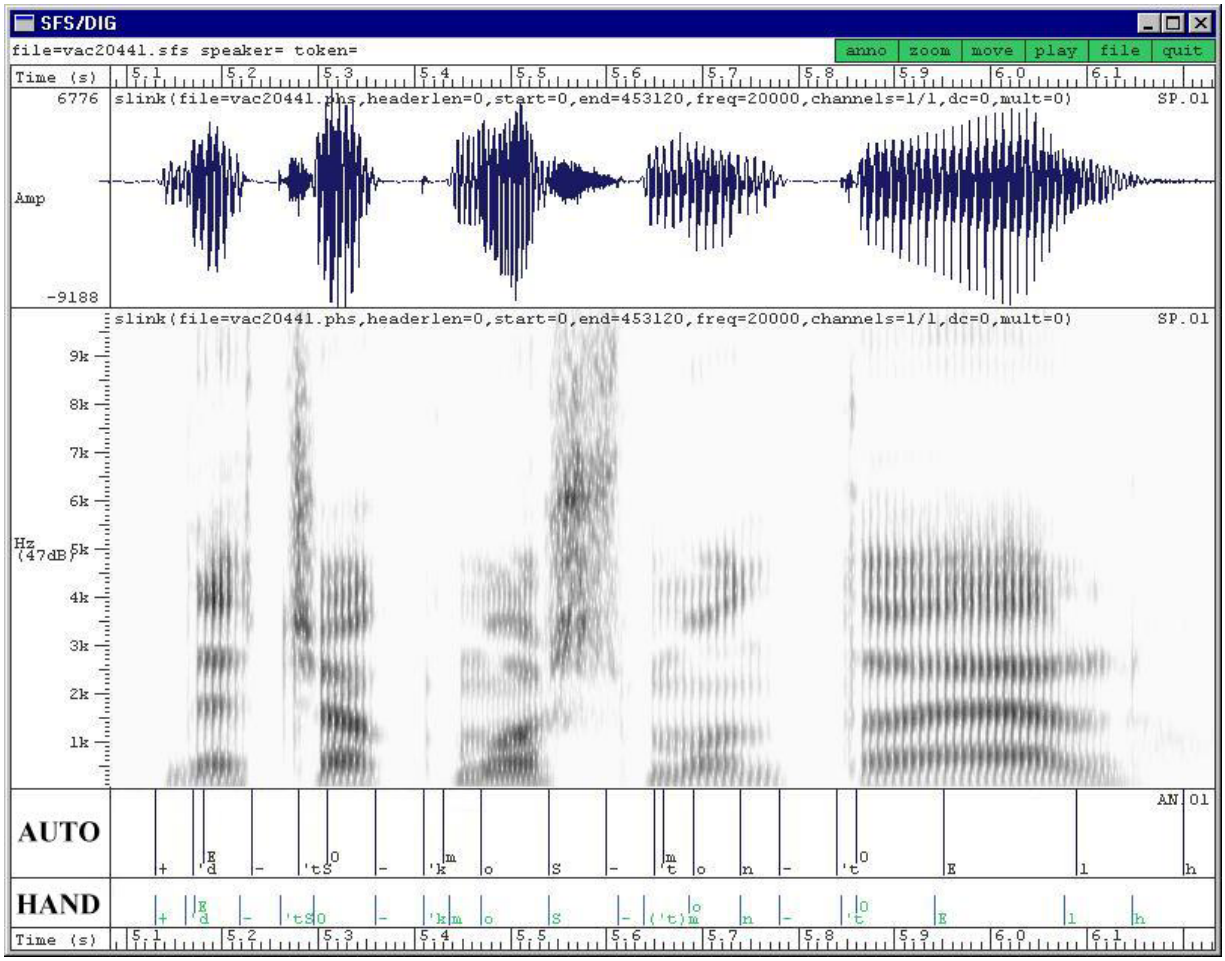
4.2.4.2 Szegmentálás

Szegmentálás során a beszéd időfüggvényében bejelölik a beszédhangok, vagy egyéb fonetikai egységek határait, és beírják a megfelelő fonetikai szimbólumokat. Kézi vagy automatikus szegmentálást szokás használni.
A kézi szegmentálás pontos, de fárasztó és időigényes. (4. ábra)



4. ábra SPECO gyermek beszédadatbázis kézi szegmentálása

Az automatikus, vagy félautomatikus szegmentálás gépi felismerésekkel hajtható végre. Gyors eljárás, de elég pontatlan, ilyenkor kézi korrekciót kell végezni.
A BME Távközlési és Telematikai Tanszékén 1998-ban kifejlesztett neurális háló alapú beszédfelismerővel végzett automatikus nyelvfüggetlen szegmentálóval készített minta az 5. ábrán látható.



5. ábra LIAS (Language Independent Automatic Segmentation Method) automatikus szegmentáló

5 Adatbázisok a gyakorlatban

Beszédadatbázisokat elsősorban a gépi beszédfelismerésben használunk. Széles felhasználói terület még az automatikus beszédszintézis, kódolás, elemzés, beszédazonosítás, nyelvazonosítás. Mindezen területek nagy adatbázist igényelnek. Nemcsak betanításra használatosak, hanem tesztelésre is, hiszen segítségükkel rögzített, állandó anyaggal, tehát ismétlési lehetőséggel tudjuk végrehajtani a tesztelést.

Célszerű lenne a beszédadatbázisok létrehozására egy egységes szabvány eljárást kidolgozni, de ennek komoly akadálya van. A felhasználási területek szélesek, a nyelvi sajátosságok különbözőek, a nemzeti érdekek erősek, így igen nehéz egységes szabványt kidolgozni (elfogadtatni?)mind az adatbázisok létrehozására, mind az értékelési eljárásokra.

Európában az ESPRIT projekt keretében szabványos beszéd és nyelvre vonatkozó rögzítési és feldolgozási eljárásokat fogadtak el Speech Assesment Methods néven (SAM), ezek rögzítési eljárások, tárolások, annotálási technikák, adatok eloszlása. E projekt keretében hozták létre a fonémák SAMPA fonetikus szimbólumrendszerét. E szabványeljárások alapján jöttek létre az EUROM 0, EUROM 1, BABEL adatbázisok Európa legtöbb nyelvét átfogva. Csendes szobai felvételeket tartalmaznak megfelelően szerkesztett hangkapcsolatokat, szavakat, mondatokat, számokat felolvasva. A SQUALE projekt keretén belül olyan adatbázist hoztak létre a kutatók, amely a beszédfelismerők értékelésére használható.

A SPEECHDAT 1,2 és a SPEECHDAT-E adatbázisok vezetékes és mobil telefonon keresztül rögzített beszédet tartalmaznak.
Ma már se szeri se száma a különböző adatbázisoknak. Néhány jellemző adatbázis adatait a 5. és 6. Táblázatban mutatjuk be. Európában az ERLA (European Resources of Language and Speech) társaság forgalmazza a legtöbb európai adatbázist, rendszeres kiadványokban tájékoztatva a szakembereket az újonnan megjelent beszéd és nyelvadatbázisokról.

5. Táblázat: Adatbázisok minőségi jellemzői

Adatbázis neve	**Forrás**	**Formátum kHz**	**Rögzítési környezet**	**Bemondás módja**	**Feldolgozás**	
					alapegység	**átírás**
TI Digits	mikrofon	20	csendes szoba	felolvasás	szó	nem
TIMIT	mikrofon	16	csendes szoba	felolvasás	beszédhang	igen
NTIMIT	telefon	8	telefonon keresztül telefonfülke, iroda, lakás, utca stb.	felolvasás	beszédhang	igen
ATISO	mikrofon	16	hivatal	spontán felolvasás	mondat	nem
Switchboard (Credit Card)	telefon	8	telefonon keresztül telefonfülke, iroda, lakás, utca stb.	spontán beszéd	szó	igen
Switchboard (Credit Card)	telefon	8	telefonon keresztül telefonfülke, iroda, lakás, utca stb.	spontán beszéd	szó	igen
MARSEC	mikrofon	16	változó	spontán	beszédhang	igen
ATIS2	mikrofon	16	hivatal	spontán	mondat	nem
EUROM-1 BABEL	mikrofon	20	csendes helyiség	olvasott	beszédhang szó	igen
SpeechDat SpeechDat-E	telefon	8	vezetékes telefon, mobil telefonfülke, iroda, lakás, utca stb.	olvasott, spontán	beszédhang szó	nem

6. Táblázat: Adatbázisok mennyiségi jellemzői

Adatbázis neve	CD száma	Felvételi idő	Méret gigabájt	Beszélők száma	Egységek száma
TI Digits	3	~14	2	3260630	>2 500 szám
TIMIT	1	5,3	0,65	630	6 300 mondat
NTIMIT	2	5,3	0,65	144	6 300 mondat
ATISO	6	20,2	2,38	69	10 722 kiejtés
Switchboard (Credit Card)	1	3,8	0,23	16	35 dialógus
Switchboard (Credit Card)	30	250	15	100	2 500 dialógus
MARSEC	1	5,5	0,62	351	53 mondat
ATIS2	6	~37	~5	>124	12 000 kiejtés
EUROM-1 BABEL	3-5	nyelvfüggő	nyelvfüggő	100	számok, fonetikailag kiegyen-súlyozott mondatok, hangkapcsolatok, szavak
SpeechDat SpeechDat-E	4-6	nyelvfüggő	nyelvfüggő 1,5-4	nyelvfüggő 500-5000	számok, nevek, intézmények, utasítások, fonetikailag gazdag mondatok

Amerikában a TIMIT és az ATIS a legjelentősebb - gépi beszédfelismerés céljára létrehozott - adatbázis, amelyek amerikai angol nyelven akusztikai modellek felépítésére alkalmasak.
A TIMIT adatbázis személyfüggetlen fonetikai beszédfelismerők betanítására és tesztelésére szolgál. Szómodellek felépítésére alkalmatlan, mivel szűkített szótárkészletet használ, fonetikailag gazdag mondatai viszont kiválóan alkalmasak beszédhang modellek létrehozására. Az adatbázis egy része betanításra, másik része tesztelésre ad lehetőséget.
ATIS (Air Travel Information System) repülőtéri információval kapcsolatos szótárkészleten alapuló adatbázis. Minden elem rögzítésre került, spontán társalgással, és olvasva hivatali körülmények között.

6 Magyar nyelvű adatbázisok

Ezideig 3 egymástól igen különböző magyar beszédadatbázis készült el, amelyek összefoglaló adatai a 7. Táblázatban láthatóak.

7. Táblázat Magyar beszédadatbázisok összefoglaló adatai

	BABEL	**SpeechDat-E**	**SPECO gyermek adatbázis**
forrás	mikrofon	telefon	mikrofon
formátum	20 kHz, 16 bit	8 kHz, 16 bit (ISDN)	20050 Hz, 16 bit
rögzítési környezet	süketszoba (tiszta beszéd)	iroda, lakás, utca, telefonfülke stb.	süketszoba
bemondás módja	olvasott szöveg	80% olvasott, 20% spontán szöveg	olvasott, utánzott szöveg
szövegtípus	hangkapcsolatok számok, szavak folyamatos szöveg	betűzött szavak dátumok, pénzösszegek számok, telefon- és hitelkártyaszámok szavak, tulajdonnevek, mondatok	kitartott beszédhangok hangkapcsolatok számok, szavak mondatok
bemondók száma	60	1000	76
feldolgozás	fonotipikus átírás fonémaszintű szegmentálás	karakteres leírás nincs szegmentálás zajok, hibák jelölése	fonotipikus átírás fonémaszintű szegmentálás

6.1 BABEL magyar nyelvű adatbázis

Az első magyar beszédadatbázis egy többnyelvű kelet-európai beszédadatbázist létrehozó munkaprogram keretében készült el 1995 és 1998 között. A munkaprogram összefoglaló neve BABEL. Célja egy közös, egységes elvek alapján felépített nagyméretű beszédadatbázis létrehozása a beszédakusztikával, fonetikával, digitális jelfeldolgozással, valamint nyelvészettel foglalkozó európai szakemberek munkájának segítésére. A BABEL program keretén belül 5 közép-és kelet-európai nyelv beszédadatbázisa készült el, ezek a nyelvek: bolgár, észt, magyar, lengyel és román.

A magyar BABEL adatbázis a hivatalos magyar köznyelvet reprezentáló rendezett hanganyag, amely hangkapcsolatokat, szavakat, számokat, 5 mondatos bekezdéseket tartalmaz, valamint 120 bekezdés fonetikai szinten címkézett és szegmentált anyagát.

Az adatbázis erősen zajcsökkentett környezetben felvett olvasott szöveg, ez az ún. tiszta olvasott beszéd, melyet 60 személlyel, 30 férfivel és 30 nővel rögzítettünk kor és foglalkozás szerint széles eloszlásban. A teljes hanganyag 1,8 GB terjedelmű, amelyet 3 CD-n rögzítettünk.

Az adatbázis összetétele

Az adatbázis összetétele és formája az ESPRIT programban kialakított SAM szabályokat követi.
Az adatbázis szövegkészlete 3 részből áll:
- rövid bekezdések, amelyek egyenként 5 tematikailag összefüggő mondatot tartalmaznak;
- kiválasztott számok 0-9999-ig;
- szisztematikusan megszerkesztett CVC hangkapcsolatok különállóan és mondatba szer kesztve.

40 különböző bekezdés folyamatosan olvasott mondatokból áll, amelyből az első rész, a 30 speciálisan megszerkesztett bekezdés, eltér az EUROM1 szabvány-előírásaitól. A szabványelőírás alapján létrehozott szöveg hangzó és szótag statisztikáját megvizsgálva, az anyag igen szegényesnek bizonyult, ugyanis nem írja le a magyar fonéma-kapcsolódások teljes készletét. Ezért e 30 bekezdést a szabványtól eltérő módon szerkesztettük meg. A magyar nyelv részletes statisztikai elemzése alapján már korábban azt találtuk, hogy a félszótag egység írja le a legtömörebben a magyar nyelv fonológiai szerkezetét. Egy általános magyar szöveg 98%-a leírható 491 félszótaggal, 99%-a pedig 600 félszótaggal. Ez annyit jelent, hogy ha a félszótag eloszlásnak megfelelő szöveget szerkesztünk meg, akkor fedjük le legtömörebben a magyar nyelv hangkapcsolat-variációit.

A beszélők kiválasztása

A jelen adatbázis, a magyar köznyelvet reprezentálja, tehát a különböző dialektusok nincsenek benne képviselve, de a magyar köznyelvi beszéd olyan széles variációit rögzítettük, amely az adott körülmények között lehetséges volt. Az olvasásnál az egyetlen kritérium az volt, hogy pontosan azt kell felolvasni, ami le van írva. Budapesten élő és dolgozó férfiak és nők voltak a beszélők, 14-69 éves kor között. A 60 beszélő kor szerinti megoszlását a 8. Táblázat mutatja be.

8. Táblázat: A beszélők nem és kor szerinti eloszlása (3 női tartalék beszélővel)

Életkor	Beszélők száma (férfiak)	Beszélők száma (nők)
14-19	3	1
20-29	8	7
30-39	6	6
40-49	5	10
50-59	6	8
60-69	2	1

Szegmentálás és címkézés
Az anyagban összesen 120 paragrafus került fonetikai szintű szegmentálásra és címkézésre. Kézi szegmentálással, a beszéd időfüggvényében, bejelöltük a fizikailag megfigyelhető fonémák határait és beírtuk a megfelelő helyre a fonéma címkéket, audio-vizuális fonetikai átírással. A fonetikai átírás a SAMPA készlet segítségével készült.

A BABEL adatbázis viszonylag kis számú bemondóval készült, viszont az összeállított szöveg viszonylag hosszú. Létrehozásánál az volt a cél, hogy jó alapanyagot teremtsen fonetikai kutatásokhoz: a hangkörnyezetnek, a hang helyzetének, a különböző szupraszegmentális jegyeknek stb. a hatása jól vizsgálható legyen. A speciális paragrafusszöveg miatt az alapfonetikai kutatások mellett a hanganyag a beszédfelismerési kutatásoknak is alapot tud biztosítani.

A felvételek a Békésy György Akusztikai Kutatólaboratórium süketszobájában készültek.

6.2 SpeechDat-E telefonbeszéd adatbázis

Az adatbázis 1000 beszélő által telefonon bemondott szövegből áll. Számos különböző típusú, telefonon keresztül működő beszédfelismerő betanítására és tesztelésére ad lehetőséget. Ezek az izolált szavas rendszerek, szókereső és azonosító rendszerek, dialógus rendszerek, valamint szótárfüggetlen rendszerek, amelyeknél a felismerés szónál kisebb felismerési egységek modellezésén alapul. Az összeállított szöveganyag a sokfeladatos elvárásoknak megfelelően igen sokrétű.
Tartalmaz: parancsszavakat, számjegysorozatokat, telefonszámot hitelkártyaszámot, PIN kódot, spontán dátumot, relatív dátumot, parancsszavas kifejezést, számjegyet, betűzött spontán vezetéknevet, betűzött városnevet, betűzött szót, pénzmennyiséget (forint/euro), természetes számot, spontán vezetéknevet, spontán városnevet, cégnevet, vezeték+keresztnevet, igen/nem kérdést igen/nem válasszal, fonetikailag gazdag mondatot.

A telefon-beszédadatbázis specifikációja az MLAP LRE-63343 SPEECHDAT (M) EU projekt javaslata alapján készült. Ez biztosítja azt, hogy a különböző nyelvű adatbázisok igen hasonlóak, egységes alapot képviselve ugyanazt a beszédtechnológia fejlesztési lehetőséget nyújtsák a feldolgozott nyelvhez.
Az adatbázis a BME Távközlési és Telematikai Tanszékén készült.

6.3 SPECO gyermekbeszéd adatbázis

Az adatbázis csendes helyiségben 5-10 éves gyermekek által bemondott szótagokat, szavakat, mondatokat tartalmaz. A fonetikai, beszédfelismerési kutatásokhoz (hangkörnyezet, hanghelyzet, különböző szupraszegmentális jegyek stb. vizsgálata) biztosít megfelelő hanganyagot

Az adatbázis nagy részében átlagos köznyelvi olvasott gyermekbeszéd van rögzítve csendes korülmények között. Kisebb részben pösze, és különböző súlyosságú hallássérült gyermekek beszédjét tartalmazza

Az adatbázis a BME Távközlési és Telematikai Tanszékén készült.