Uporaba kanoničnega govornega akustičnega modela za prilagajanje prostora govornih akustičnih značilk

Simon Dobrišek*, Boštjan Vesnicer*, Jerneja Žganec Gros†, France Mihelič*

*LUKS, Fakulteta za elektrotehniko, Univerza v Ljubljani
Tržaška 25, 1000 Ljubljana
simon.dobrisek@fe.uni-lj.si

†Alpineon, d.o.o
Ulica Iga Grudna 15, 1000 Ljubljana

Povzetek

V članku predstavljamo rezultate poskusov s postopki prilagajanja akustičnega govornega modela na globalne akustične značilnosti govornih sej. Preizkušeni postopki temeljijo na CMLLR-transformacijah. Uporabili smo poseben, manj obsežen kanonični govorni akustični model, katerega namen je izključno določanje CMLLR-transformacij. Parametre bolj obsežnega govornega akustičnega modela smo določali po posebnem učnem načrtu z uporabo že transformiranih vektorjev akustičnih značilk. Rezultati preizkusov samodejnega razpoznavalnika govora z govornimi sejami, ki po globalnih akustičnih in govornih značilnostih odstopajo od učnih govornih sej, so v primerjavi z izhodiščnim modelom, ki ne izvaja opisanega postopka prilagajanja, pokazali izboljšanje pravilnosti razpoznavanja govora.

Adaptation of Acoustic Feature Space Using Canonical Acoustic Model

The paper presents the results of experiments with a speaker-adaptive-training scheme that is based on CMLLR. A simple canonical acoustic model was used to obtain linear transformations of acoustic feature space that are speech session dependent. A more complex acoustic model, used for the actual automatic speech recognition, was first initialized from the canonical model. Its parameters were then reestimated from the acoustic feature vectors that were previously transformed using the session-dependent linear transformations. The presented results indicate that, when test speech sessions differ considerably from the training ones, automatic speech recognition is improved using the proposed training scheme.

1. Uvod

Trenutno najboljši samodejni razpoznavalniki govora podpirajo možnost samodejnega prilagajanja akustičnega govornega modela akustičnim značilnostim trenutne govorne seje. Govorna seja zajema vse govorne posnetke istega govorca, ki so posneti v približno enakih akustičnih razmerah. Pogosto so vsi posnetki istega govorca obravnavani kot ena govorna seja.

Uveljavljeni postopki tovrstnega prilagajanja praviloma predpostavljajo uporabo akustičnih govornih modelov, ki temeljijo na teoriji prikritih Markovovih modelov (PMM). V preteklih letih je bilo narejenega največ dela predvsem na postopkih prilagajanja z uporabo linearnih transformacij akustičnih govornih modelov (Gales, 1998; in P. C. Woodland, 2001). Poleg te možnosti so najbolj znani še postopki, ki temeljijo na kriteriju največjega aposteriornega verjetja modela (angl. MAP) (in C. H. Lee, 1994). Na slednjih je bilo v zadnjem času opravljenega veliko dela predvsem pri razvoju samodejnih razpoznavalnikov govorcev.

Že pred leti smo si za raziskovalni cilj zastavili razvoj lastnega pogona za samodejno razpoznavanje tekočega slovenskega govora z velikim besednjakom, ki bo imel zmožnost samodejnega prilagajanja na govorne seje in bo primeren za uporabo v vgradnih sistemih. V zadnjem času smo delali predvsem na postopkih, ki temeljijo na linearnih transformacijah in omogočajo sprotno prilagajanje govornega modela trenutnim govornim sejam.

Posebej smo se posvetili določanju omejenih globalnih linearnih transformacij parametrov akustičnega govornega modela, ki jih je mogoče preslikati v linearne transformacije prostora akustičnih govornih značilk. Z določanjem takšnih globalnih transformacij, ki so od govornih sej odvisne, lahko namreč zgradimo kanonični akustični govorni model, ki je deloma neodvisen od globalnih akustičnih značilnosti govornih sej. Transformacije, dobljene s kanoničnim modelom, lahko nato uporabimo pri učenju običajnega samodejnega razpoznavalnika govora, ki tako tudi postane deloma neodvisen od globalnih akustičnih značilnosti govornih sej. To je eden od možnih načrtov postopka učenja, ki se prilagaja govornim sejam in s tem tudi govorcu (angl. Speaker Adaptive Training - SAT).

Članek opisuje nekaj naših poskusov z različnimi načrti postopka učenja razpoznavalnika govora, ki se prilagaja govornim sejam na prej opisan način. Zaradi časovne zahtevnosti izvajanja takšnih poskusov smo primerjali rezultate, dosežene z samodejnim razpoznavalnikom s srednje velikim besednjakom. Za izvajanje poskusov smo v pretežni meri uporabljali orodje HTK. Za bolj učinkovito veriženje linearnih transformacij, ki so posledica iterativnega postopka ocenjevanja parametrov, smo razvili tudi nekaj lastnih orodij.

2. Prilagajanje z linearnimi transformacijami

Pri akustičnih govornih modelih, ki temeljijo na teoriji PMM, se prilagajanje z linearnimi transformacijami ne nanaša na prav vse parametre tega modela. Ponavadi se izvaja linearno transformacijo le na parametrih funkcij normalnih gostot verjetnosti, s katerimi modeliramo porazdelitve naključnih spremenljivk v posameznih stanjih naključnega avtomata. V tem primeru se transformacije nanašajo le na srednje vrednosti in variance Gaussovih porazdelitev. Prehodne verjetnosti med stanji naključnega avtomata in apriorne verjetnosti Gaussovih komponent v mešanicah, ki modelirajo omenjene porazdelitve, pa v teh postopkih prilagajanja ponavadi ne spreminjamo.

Obstaja več vrst linearnih transformacij, ki se uporabljajo za prilagajanje akustičnega govornega modela (in M. J. F. Gales, 2005). Mi smo se posvetili predvsem linearnim transformacijam, ki jih je mogoče preslikati iz transformacije parametrov akustičnega govornega modela v transformacijo prostora akustičnih govornih značilk. Primer takšne transformacije je omejena linearna transformacija, določena po kriteriju največjega verjetja akustičnega modela (angl. Constrained Maximum Likelihood Linear Regression - CMLLR) (Gales, 1998).

Pri CMLLR-transformaciji se vektorji srednjih vrednosti μ in kovariančne matrike $\mathbf{\Sigma}$ linearno transformirajo po spodnjih enačbah.

$$\hat{\mu} = \mathbf{A}'\mu - \mathbf{b}' \quad , \quad \hat{\mathbf{\Sigma}} = \mathbf{A}'\mathbf{\Sigma}\mathbf{A}'^{\mathbf{T}}$$

Matrika $\mathbf{A}'$ in vektor $\mathbf{b}'$ predstavljata linearno transformacijo in njune koeficiente določamo po kriteriju največjega verjetja akustičnega modela za dane nize vektorjev govornih akustičnih značilk $\mathbf{o}(\tau)$, ki so na razpolaga za prilagajanje. Določanje koeficientov matrike $\mathbf{A}'$ in vektorja $\mathbf{b}'$ izvedemo s uveljavljenim postopkom EM (angl. Expectation-Maximization) kot je podano v (Gales, 1998).

Dobljeno transformacijo parametrov Gaussovih porazdelitev lahko enostavno preslikamo v linearno transformacijo vektorjev akustičnih značilk, kot je podano v spodnjem izrazu.

$$\hat{\mathbf{o}}(\tau) = \mathbf{A}\mathbf{o}(\tau) + \mathbf{b} = \mathbf{A}'^{-\mathbf{1}}\mathbf{o}(\tau) + \mathbf{A}'^{-\mathbf{1}}\mathbf{b}'$$

Matriko $\mathbf{A}$ in vektor $\mathbf{b}$ ponavadi združimo v matriko $\mathbf{W} = [\mathbf{A}\ \mathbf{b}]$, ki nato enovito predstavlja iskano linearno transformacijo. Koeficiente matrike $\mathbf{W}$ določamo iz govornih posnetkov za vsako govorno sejo posebej. S tem pridemo do linearnih transformacij $\mathbf{W}(s)$, ki vektorje akustičnih značilk dane govorne seje s prilagodijo akustičnemu govornemu modelu.

3. Načrt učenja s prilagajanjem

Zahteva po sprotnem prilagajanju akustičnega govornega modela trenutni govorni seji ponavadi pomeni, da je za ta namen na razpolago razmeroma malo govora. Zaradi statistične narave akustičnega govornega modela in postopka ocenjevanja koeficientov linearne transformacije je očitno, da je pri majhni količini posnetkov prilagajanje boljše, če ima model manjše število parametrov. Zato smo se odločili, da bomo za prilagajanje uporabili posebni akustični govorni model z manjšim številom parametrov. Namen tega osnovnega modela je bil izključno ocenjevanje koeficientov linearne transformacije, ki se je uporabljala za sprotno transformiranje vektorjev akustičnih značilk. Dejanski razpoznavalnik govora smo učili in ga preizkušali na že transformiranih vektorjih akustičnih značilk.

Osnovni akustični govorni model smo sestavili kot razpoznavalnik kontekstno neodvisnih alofonov. Za vsakega od dvaintrideset alofonov ter treh dodatnih akustičnih enot (tišina, tlesk z jezikom in vdih) smo uporabil običajne levo-desne PMM s tremi stanji. Ta model smo uporabili tudi kot izhodišče pri določanju in učenju znatno večjega števila akustičnih modelov kontekstno odvisnih alofonov - trifonov.

Obravnavan učni načrt je možen le v primeru, ko so učni govorni posnetki urejeni po govornih sejah. Za vsako govorno sejo s smo določali njej lastno matriko $\mathbf{W}(s)$. Začetne vrednosti koeficientov matrik $\mathbf{W}(s)$ smo inicializirali tako, da smo normalizirali globalne vektorje srednjih vrednosti $\hat{\mu_0}(s) = \mathbf{0}$ in kovariančne matrike $\hat{\mathbf{\Sigma}_0}(s) = \mathbf{I}$, ki sta ocenjena iz vseh transformiranih $\hat{\mathbf{o}}(\tau)$ dane govorne seje s. To se enostavno doseže tako, da se koeficiente matrike $\mathbf{W}(s) = [\mathbf{A}(s)\ \mathbf{b}(s)]$ inicializira na sledeč način

$$\mathbf{A(s)} = \mathbf{L}_0(s)^T \ , \ \mathbf{b}(s) = -\mathbf{L}_0(s)^T\mu_0(s) \ ,$$

kjer $\mu_0(s)$ označuje globalni vektor srednjih vrednosti in $\mathbf{L}(s)$ spodnjo trikotno matriko razcepa Choleskega inverzne globalne kovariančne matrike $\mathbf{\Sigma}_0(s)^{-1}$. Pri tem sta $\mu_0(s)$ in $\mathbf{\Sigma}_0(s)$ ocenjena iz netransformiranih $\mathbf{o}(\tau)$ dane govorne seje s.

Takšno inicializacijo smo izvedli zato, ker smo smo za akustične značilke uporabili običajne MFCC-koeficiente. Na ta način smo v prilagajanje z linearnimi transformacijami vključili še normalizacijo MFCC-koeficientov po srednjih vrednostih in kovariancah (angl. Cepstral Mean and Covariance Normalization). Za takšno normalizacijo je znano, da zmanjšuje občutljivost govornega modela na akustično spremenljivost govornih sej.

Učenje s prilagajanjem smo v grobem izvajali po naslednjih korakih:

- Izračun globalnih vektorjev srednjih vrednosti $\mu_0(s)$ in kovariančnih matrik $\mathbf{\Sigma}_0(s)$ za vsako učno govorno sejo s.
- Inicializacija matrik $\mathbf{W}(s)$ za vsako učno govorno sejo s z uporabo prej izračunanih $\mu_0(s)$ in $\mathbf{\Sigma}_0(s)$.
- Izmenično iterativno ocenjevanje novih vrednosti matrik $\mathbf{W}(s)$ in parametrov osnovnih kanoničnih akustičnih govornih modelov iz vseh učnih govornih sej.
- Inicializacija parametrov akustičnih modelov trifonov z uporabo parametrov osnovnih akustičnih govornih modelov alofonov.
- Iterativno ocenjevanje parametrov akustičnih modelov trifonov iz vseh učnih govornih sej s z upoštevanjem linearnih transformacij, ki jih določajo matrike $\mathbf{W}(s)$.

Pri osnovni različici učnega načrta smo matrike $\mathbf{W}(s)$ inicializirali na običajen način z enotsko matriko in ničelnim vektorjem.

3.1. Govorne zbirke

V učno govorno zbirko smo združili tri različne zbirke. Zbirki Gopolis in K211d vsebujeta pretežno bran govor, posnet v nadzorovanem akustičnem okolju s kakovostnim mikrofonom. Zbirka VNTV pa vsebuje običajne televizijske posnetke vremenskih napovedi, ki so jih voditelji podali v okviru dnevnih poročil na Televiziji Slovenija. Učna govorna zbirka je tako vsebovala posnetke govornih sej petinšestdesetih govorcev. Skupno trajanje vseh učnih posnetkov je približno dvanajst ur in pol. Za eno govorno sejo smo šteli vse posnetke istega govorca. Iz učne govorne zbirke smo izločili tristo posnetkov, ki smo jih namenili za preizkus samodejnega razpoznavanja, ki je bil od učnih govornih sej odvisen.

Preizkusne govorne posnetke, ki so od učnih govornih sej deloma neodvisni, smo pridobili posebej za izvedbo poskusov, opisanih v tem članku. Dvaindvajset govorcev (v glavnem študentov) smo prosili, da posnamejo po dvajset daljših stavkov. Naključno tvorjeni stavki so se nanašali na poizvedovanja po letalskih informacijah. Do teh stavkov smo prišli podobno kot pri pridobivanju zbirke Gopolis (S. Dobrišek, 1998). Testni posnetki so bili pridobljeni v nenadzorovanih akustičnih okoljih in z različnimi mikrofoni, računalniki ter programi za snemanje zvoka. Pri testnih posnetkih gre še vedno pretežno za bran govor, a se ta govor znatno razlikuje od posnetkov v zbirki Gopolis. Govorcev namreč nismo posebej motivirali, da bi stavke jasno artikulirali, zato se pri znatnem številu posnetkov odražajo prvine spontanega govora (tleskanje z jezikom, vzdihi ipd).

4. Zgradba govornih modelov

Pri izvedbi poskusov smo poskrbeli za čim večjo primerljivost med preizkušenimi govornimi modeli. Vsi govorni modeli so bili tvorjeni s pomočjo orodij iz zbirke HTK. Orodjem smo dodali le možnost bolj učinkovitega veriženja linearnih transformacij. V vseh poskusih smo uporabljali iste govorne zbirke, vektorje akustični značilk in govorne modele z istim številom parametrov. Slednje postane pomembno predvsem pri izvedbi vezave parametrov s fonetičnimi odločitvenimi drevesi. Pri tem postopku je končno število parametrov govornega akustičnega modela odvisno od določenega praga (S. Young, 2005). Za povsem enako število parametrov smo poskrbeli tako, da smo pri določanju praga in doseganju želenega števila parametrov uporabljali rekurzivni postopek bisekcije.

Za vektorje akustičnih značilk smo uporabljali običajne 39-razsežne vektorje, sestavljene iz MFCC-koeficientov in njihovih delta- in delta-delta koeficientov. Iskane linearne transformacije so se nanašale na celotne 39-razsežne vektorje.

Pri vseh akustičnih modelih smo uporabil običajne levo-desne PMM s tremi stanji. Osnovni kanonični akustični model je tako poleg verjetnosti prehodov med stanji PMM tvorilo še 105 Gaussovih funkcij gostot verjetnosti z diagonalnimi kovariančnimi matrikami. Trifonski akustični modeli so imeli po alofonih vezane verjetnosti prehodov med stanji PMM in 3200 vezanih stanj s po pet-komponentnimi Gaussovimi porazdelitvami. Ta akustični model je tako vseboval skupaj točno 16000 Gaussovih funkcij gostot verjetnosti z diagonalnimi kovariančnimi matrikami. Fonetična vprašanja, ki so potrebna za vezavo parametrov smo tvorili ročno (Dobrišek, 2001) in v kombinaciji z vprašanji, samodejno pridobljenimi z orodji, ki so del zbirke Sphinx III.

Poleg navedenih parametrov imajo na rezultat razpoznavanja precejšen vpliv tudi drugi parametri Viterbijevega postopka iskanja najbolj verjetnega zaporedja stanj govornega modela pri danem govornem posnetku. Pri teh parametrih smo pazili predvsem na to, da smo pri vseh govornih modelih dosegli približno enak čas razpoznavanja istih govornih posnetkov. Vedno smo tudi uporabljali enako razmerje med vplivom akustičnega in jezikovnega modela na rezultat razpoznavanja in poskrbeli za približno enako razmerje med napakami vrivanj in izbrisov govornih enot.

4.1. Preizkušanje razpoznavalnikov

Vse zgrajene razpoznavalnike smo preizkušali z ugotavljanjem napak pri samodejnem razpoznavanju glasov in besed. Pri razpoznavanju glasov (alofonov) nismo uporabljali nobenega jezikovnega modela. To pomeni, da je govorni model vključeval predpostavko, da vsak alofon lahko sledi drugemu z enako verjetnostjo. Pri razpoznavanju besed smo upoštevali besednjak s približno pettisoč besedami. Govorni model je vključeval bigramski jezikovni model, ocenjen iz učne govorne zbirke. To pomeni, da je vključeval tako poizvedovanja po letalskih informacijah kot tudi vremenske napovedi. Kot smo že omenili, so se preizkusne govorne seje nanašale le na poizvedovanja po letalskih informacijah. Preizkusni govorci niso bili vključeni v učno govorno zbirko.

Pri preizkušanju razpoznavalnikov, ki se prilagajajo na nove govorne seje, se pojavi problem začetne ocene linearnih transformacij, ki prilagodijo govorni model njihovim globalnim akustičnim značilnostim. Preizkus smo si zaenkrat zamislili tako, da smo del preizkusnih posnetkov namenili izključno začetnemu prilagajanju govornega modela in nato uporabili preostali del za dejanski preizkus pravilnosti razpoznavanja. S poskusi smo ugotovili, da se doseže dobre rezultate že z desetimi poljubnimi krajšimi stavki, ki se namenijo izključno začetnemu nenadzorovanemu prilagajanju govornega modela.

V praksi bi to pomenilo, da bi moral nov govorec najprej izgovoriti deset poljubnih stavkov, ki bi bili namenjeni izključno začetnemu prilagajanju govornega modela na njegove globalne akustične značilnosti. Rezultati, ki so podani v tem članku, predpostavljajo takšno začetno prilagajanje. Nadaljnje prilagajanje se je nato izvajalo sprotno z vsakim novim preizkusnim stavkom, ki ga je izgovoril govorec. V naših prihodnjih poskusih nameravamo oceniti tudi kako narašča pravilnost razpoznavanja od prvega stavka naprej. Ta podatek je zanimiv za primere, ko govorcev ne bi radi obremenjevali s takšnim začetnim prilagajanjem govornega modela.

5. Rezultati

Podajamo rezultate razpoznavanj za štiri vrste poskusov. Pri prvem poskusu (BASE) je bil uporabljen model, ki se ni prilagajal na globalne akustične značilnosti govornih sej. Pri drugem poskusu (CVMN) smo uporabljali linearne transformacije, s katerimi izvedemo le normalizacijo srednji vrednosti in varianc globalnih Gaussovih porazdelitev posameznih govorih sej. Pri tretjem poskusu (CMLLR) smo izvedli prilagajanje na opisan način s kanoničnim akustičnim modelom, pri katerem so bile transformacije inicializirane na običajen način z enotsko matriko in ničelnim vektorjem. Pri zadnjem poskusu (CVMN-CMLLR) pa smo transformacije inicializirali iz srednji vrednosti in varianc globalnih Gaussovih porazdelitev posameznih govorih sej. Rezultati preizkusnih razpoznavanj glasov (alofonov) so

MODEL	EVAL	ADPT	TEST
BASE	89,7%	57,2%	56,8%
CVMN	90,1%	59,8%	59,1%
CMLLR	92,0%	62,2%	62,1%
CVMN-CMLLR	92,4%	63,4%	63,2%

Tabela 1: Ocenjene verjetnosti pravilnega razpoznavanja glasov pri različni govornih akustičnih modelih

podani v tabeli 1. Rezultati so podani kot ocene verjetnosti pravilnega razpoznavanja glasov. Rezultati, označeni z EVAL, se nanašajo na že omenjenih tristo posnetkov, ki so bili naključno izbrani in izločeni iz učne govorne zbirke. Ti rezultati predstavljajo oceno verjetnosti pravilnega razpoznavanja glasov v posnetkih, ki so od učnih govornih sej odvisni. Rezultati, označeni z ADPT, se nanašajo na preizkusne posnetke, ki so bili uporabljeni za začetno oceno linearnih transformacij pri prilagajanju govornega modela. Rezultati, označeni z TEST, pa se nanašajo na dejanske preizkusne posnetke, pri katerih se je izvajalo sprotno prilagajanje govornega modela. Pri rezultatih v tabeli 1 je najbolj

MODEL	EVAL	ADPT	TEST
BASE	8,7%	24,7%	26,2%
CVMN	8,8%	23,7%	25,6%
CMLLR	8,5%	19,1%	19,2%
CVMN-CMLLR	8,3%	18,8%	18,9%

Tabela 2: Ocenjene verjetnosti napačnega razpoznavanja besed pri različni govornih akustičnih modelih

opazna znatna razlika med oceno pravilnosti razpoznavanja posnetkov, ki so od učnih govornih sej odvisni, v primerjavi s tistimi, ki so od učnih govornih sej neodvisni. Glede na razmeroma velik obseg učne govorne zbirke to priča o tem, da se preizkusne govorne seje po globalnih akustičnih značilnostih res precej razlikujejo od učnih govornih sej. Po drugi strani pa je razlika med ocenami verjetnosti pravilnega razpoznavanja glasov v posnetkih ADPT in TEST majhna. To priča o tem, da po globalnih akustičnih značilnostih ni znatnih razlik med posnetki iste govorne seje, torej posnetki, ki so bili namenjeni začetni oceni linearnih transformacij in posnetki, na katerih se je izvajalo sprotno prilagajanje in dejanski preizkus razpoznavalnika.

Rezultati preizkusov razpoznavanj besed so podani v tabeli 2. Rezultati so podani kot ocene verjetnosti napačnega razpoznavanja besed. Tu so razlike med rezultati po različnih skupinah posnetkov nekaj manjši. To priča o znatnem vplivu jezikovnega modela na končni rezultat preizkusov.

6. Zaključek

Rezultatov naših poskusov potrjujejo domnevo, da postopki prilagajanja govornih modelov na globalne akustične značilnosti govornih sej z uporabo linearnih transformacij izboljšajo pravilnost samodejnega razpoznavanja govornih enot. Rezultati kažejo tudi na to, da je smiselno inicializirati linearne tranformacije tako, da v izhodišču dosežemo normalizacijo srednjih vrednosti in varianc globalnih Gaussovih porazdelitev posameznih govorih sej.

V naših nadaljnjih poskusih s postopki prilagajanja govornih modelov na globalne akustične značilnosti govornih sej se bomo posvetili predvsem postopkom, ki temeljijo na kriteriju največjega aposteriornega verjetja modela. Tudi v tem primeru bomo poskušali priti do kanoničnega govornega modela, pri katerih bo mogoče transformacije modela preslikati v transformacije govornega akustičnega prostora oziroma vektorjev govornih akustičnih značilk. Pri tem se bomo naslanjali na izkušnje, ki smo jih pridobili pri razvoju sistemov za samodejno razpoznavanje govorcev.

7. Literatura

S. Dobrišek. 2001. *Analiza in razpoznavanje glasov v govornem signalu*. Doktorska disertacija, Univerza v Ljubljani, Fakulteta za elektrotehniko.

M. J. F. Gales. 1998. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech and Language*, 12, 75–98.

J. L. Gauvain in C. H. Lee. 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. Speech Audio Proc.*, 2, 291–298.

H. Liao in M. J. F. Gales. 2005. Joint uncertainty decoding for noise robust speech recognition. V: *INTERSPEECH-2005*, str. 3129–3132.

L. F. Uebel in P. C. Woodland. 2001. Improvements in linear transforms based speaker adaptation. V: *ICASSP-2001*, str. 3129–3132.

F. Mihelič in N. Pavešič S. Dobrišek, J. Ž. Gros. 1998. Recording and labelling of the gopolis slovenian speech database. V: A. Rubio, ur., *First International Conference on Language Resources & Evaluation: Proceedings*, str. 1089–1096. European Language Resources Association.

M. Gales. T. Hain D. Kershaw G. Moore J. Odell D. Ollason D. Povey V. Valtchev in P. Woodland S. Young, G. Evermann. 2005. *The HTK Book (for HTK Version 3.3)*. Cambridge University, Engineering Department, Cambridge.