AKUSTICKÉ LISTY

České akustické společnosti

ročník 8, číslo 3 září 2002

Obsah

ČESKÁ AKUSTICKÁ SPOLEČNOST

Vážené kolegyně a kolegové,

rád bych se několika řádky zmínil o akci, která možná zůstala trochu ve stínu vašich pracovních aktivit, ale účast na ní rozhodně stála za to. Jedná se o 32. mezinárodní akustickou konferenci, která se konala ve dnech 10. – 12. září 2002 v Banské Štiavnici ve spolupráci Slovenské akustické společnosti, EAA, ČsAS, Technické univerzity ve Zvolenu a dalšími. Vzhledem k tomu, že na přípravě této konference se podíleli i členové ČSAS předpokládám, že se v příštích Akustických listech objeví ucelenější informace.

Nosným tématem konference byla sice hudební akustika, ale byly předneseny i příspěvky, které s hudební akustikou souvisely pouze okrajově. Rozsah příspěvků byl od laryngologie, přes hluk ve školním prostředí až po modální analýzu některých hudebních nástrojů. Tento rozsah dával každému účastníku konference dostatečnou možnost výběru. Všechny příspěvky byly na vysoké odborné úrovni a předneseny v angličtině. Přitom angličtina jako dorozumívací jazyk byla při vysoké účasti ze zahraničí (Dánsko, Německo, Francie a další) nutností.

Posun termínu konání 65. akustického semináře ve Skalském Dvoře na základě dohody mezi ČsAS a SAS se v tomto případě ukázal jako velmi účelný, neboť tím byla umožněna účast mnoha akustikům na obou akcích. Bohužel se však s velkou částí účastníků 35. mezinárodní akustické konference málokdy setkáváme na pravidelných Akustických seminářích pořádaných ČsAS. Přitom jejich účast, a i třeba trochu vybočující příspěvek z hlavního tématu semináře, by výrazně rozšířila akustické poznatky účastníků seminářů a členů společnosti. Lze si tedy jen přát, aby se počet účastníků seminářů ČsAS rozšířil i o odborníky z této oblasti a to nejen z ČR, ale i ze Slovenska.

Vilém Kunzl

Prof. Ing. Dr. Jiří Tichý, CSc.

4. srpna letošního roku oslavil profesor Jiří Tichý 75. narozeniny. Narodil se v Bratislavě a v roce 1945 začal studovat na Vysoké škole strojního a elektrotechnického inženýrství v Praze. Absolvoval v roce 1950, ale již od roku 1947 byl pomocnou vědeckou silou ve fyzikálním ústavu a později katedře fyziky, na které působil až do svého odchodu do Spojených států amerických. Během této doby získal doktorát technických věd (1953), titul kandidáta věd a v roce 1965 se habilitoval jako docent. Již za svého pobytu na katedře publikoval řadu odborných statí a učebních textů. Věnoval se zejména otázkám zvukové pohltivosti a problémům prostorové akustiky.

V USA se stal po roce 1968 profesorem na katedře architektury Pennsylvánské univerzity, kde působil do roku 1975. O roku 1975 až do léta 1997 vedl Graduate Program in Acoustics, který zahrnuje několik desítek profesorů nabízejících více jak sto kurzů z celé oblasti akustiky a souvisejících oborů. Počet absolvujících studentů (Ph.D, M.S. a M. Eng.) přesáhl v posledních letech stovku.

Působil a působí v celé řadě prestižních vědeckých společnostech nejen ve Spojených státech. V letech 1986 – 1987 byl přesedou společnosti Institut of Noise Control Engineering. Pracoval v oblasti mezinárodní normalizace: ISO/TC43 – Akustika, IEC/TC 29 – Elektroakustika a Americké společnosti pro normalizaci (ANSI). Americkou akustickou společností byl v roce 1991 zvolen místopředsedou společnosti a v roce 1993 předsedou společnosti. Tuto náročnou funkci zastával až do roku 1995. Organizoval a předsedal mnoha národním a mezinárodním konferencím a seminářům po celém světě (Švedsko, Francie, Japonsko, Singapůr, Brazilie a samozřejmě USA). V roce 1998 byl jmenován čestným členem České akustické společnosti.

Ve své vědecké práci ve Spojených státech navázal na své výsledky z Československa a na katedře architektury Pennsylvánské státní university, kde se věnoval stavební a prostorové akustice. Po roce 1975 se zaměřil na otázky snižování hluku strojů. Pracoval na vývoji a použití akustické intenzity. Více jak patnáct let je jeho hlavním odborným zájmem aktivní snižování hluku a další aplikace aktivních metod v akustice.

Výsledky svých prací prezentoval ve více než 30 vyzvaných přednáškách a 60 odborných referátech na národních a mezinárodních konferencích. I další publikační činnost profesora Tichého je velmi rozsáhlá. Je autorem více než 70 zásadních článků, spoluautorem 7 knih. Více než deset výzkumných zpráv souvisí s jeho konzultační činnosti pro takové firmy jako jsou například Magnavox, Ford Motor, IBM, Applied Acoustics Research Corporation apod.

Rada České akustické společnosti svým jménem a jménem všech členů společnosti přeje profesoru Jiřímu Tichému hodně úspěchů, zdraví a pohody a věří, že bude i nadále stejně aktivní jako dosud.

Porovnání časové odezvy signálu z reproduktoru elektrodynamického a elektrostatického

Josef Merhaut

Dvouletky 341, 100 00 Praha 10
email: jos.merhaut@worldonline.cz

In the paper is analyticaly shown that the transient distorsion of a harmonic signal by the newly designed loudspeaker based on the electrostatic principle is much lower, than that of the usual electrodynamic one. The author is using for each loadspeaker type the harmonic signals of frequencies 800 Hz or 5000 Hz, passed through the time window with the exponential onset and cessation.

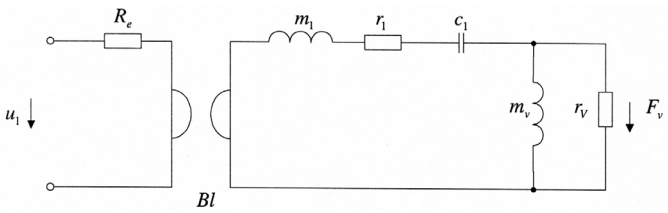
1. Úvod

V práci je proveden analytický rozbor velikosti transientního skreslení u dvou různých typů reproduktorů. Ukazuje, že z principielních důvodů je toto zkreslení u elektrostatického reproduktoru značně nižší, než u reproduktorů elektrodynamických, obvykle používaných.

K tomu účelu byly zvoleny reproduktory s obdobným frekvenčním rozsahem a podobnou velikostí: Elektrodynamický reproduktor TVM typu ARV–089–00/4 a jeho transientní odezva byla porovnána s toutéž odezvou elektrostatického reproduktoru, který byl vyvinut v rámci prací na níže zmíněném grantovém projektu. U obou reproduktorů bylo použito totéž časové okno s harmonickými signály stejných frekvencí.

2. Elektrodynamický reproduktor

Analogické schéma elektrodynamického reproduktoru v rovinné ozvučnici, se zadní stranou uzavřenou dutinou je na obrázku 1. Velikosti hodnot jednotlivých prvků (S.I.) jsou při tom u zvoleného reproduktoru tyto: hmotnost aktivní části membrány a kmitací cívky $m_1 = 8,94.10^{-4}$, poddajnost membrány včetně dutiny ozvučnice $c_1 = 2,218.10^{-5}$, mechanický odpor okraje membrány $r_1 = 1,0946$, elektrický odpor kmitací cívky $R_e = 3,06$.



Obrázek 1: Analogické schéma elektrodynamického reproduktoru

Z poloměru jeho membrány $R = 0,034$ byly stanoveny [1] tyto prvky vyzařovací impedance:

$$r_v = c_0\rho\pi R^2 = 410\pi R^2 = 1,489 \tag{1}$$

$$m_v = \rho\frac{\pi R^3}{\sqrt{2}} = 0,0001048. \tag{2}$$

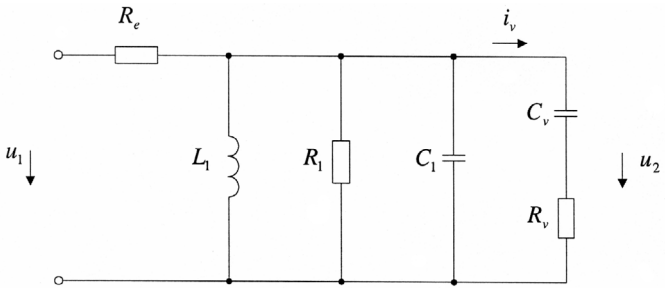
Pro vyzařovací impedanci reproduktoru platí (v uzavřené ozvučnici) [1]

$$\frac{1}{Z_v} = \frac{1}{r_v} + \frac{1}{\mathrm{j}\omega m_v} \tag{3}$$

a pro mechanickou impedanci mechanické části

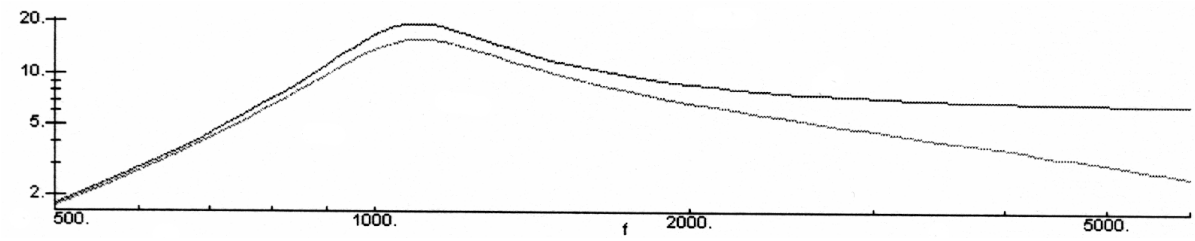
$$Z_c = \mathrm{j}\omega m_1 + r_1 + \frac{1}{\mathrm{j}\omega c_1} + Z_v. \tag{4}$$

Rezonanční frekvence je $f_r = 1130$ Hz, kritická frekvence $f_k = 2262$ Hz a převodní činitel gyrátoru $Bl = 1,873$.



Obrázek 2: Náhradní elektrické schéma elektrodynamického reproduktoru

Po převedení mechanických prvků gyrátorem na elektrickou stranu, dostaneme elektrické schéma uvedené na



Obrázek 3: Frekvenční charakteristika elektrodynamického reproduktoru. Horní křivka značí průběh pro $f_k \to \infty$

obrázku 2, ve kterém pro jednotlivé elementy dostaneme [1] výrazy

$$R_1 = \frac{(Bl)^2}{r_1} \quad (5)$$

$$C_1 = \frac{m_1}{(Bl)^2} \quad (6)$$

$$L_1 = c_1 (Bl)^2 \quad (7)$$

$$R_v = \frac{(Bl)^2}{r_v} \quad (8)$$

$$C_v = \frac{m_v}{(Bl)^2} \quad (9)$$

$$i_v Bl = F_v \quad (10)$$

Pro frekvenční přenos soustavy platí:

$$K_f = \frac{F_v}{u_1} = \frac{i_v}{u_1} Bl. \quad (11)$$

Zavedeme dále impedanci Z_x paralelního spojení dvoupólu C_v–R_v s elementy L_1, R_1 a C_1. Pro tuto impedanci platí:

$$\frac{1}{Z_x} = \frac{1}{\mathrm{j}\omega L_1} + \frac{1}{R_1} + \mathrm{j}\omega C_1 + V \frac{1}{R_v + \dfrac{1}{\mathrm{j}\omega C_v}} \quad (12)$$

Proud i_v v obrázku 2 je dán výrazem:

$$i_v = \frac{u_2}{R_v + \dfrac{1}{\mathrm{j}\omega C_v}}, \quad (13)$$

kde

$$u_2 = u_1 \frac{Z_x}{Z_x + R_e}. \quad (14)$$

Pak s použitím rovnice (11) lze psát pro frekvenční přenos

$$K_{1f} = \frac{Bli_v}{u_1} = Bl \frac{1}{1 + \dfrac{R_e}{Z_x}} \frac{1}{R_v + \dfrac{1}{\mathrm{j}\omega C_v}}. \quad (15)$$

Z toho dostaneme po úpravě konečný vztah pro přenos reproduktoru:

$$K_{1f} = \frac{\dfrac{Bl}{R_v + \dfrac{1}{\mathrm{j}\omega C_v}}}{1 + \dfrac{R_e}{R_1} + \dfrac{R_e}{\mathrm{j}\omega L_1} + \mathrm{j}\omega C_1 R_e + \dfrac{R_e}{R_v + \dfrac{1}{\mathrm{j}\omega C_v}}}. \quad (16)$$

Absolutní hodnotou výrazu (16) je modul přenosové frekvenční charakteristiky měniče. Je znázorněna v obrázku 3. Na tomto obrázku je také zobrazena pro další použití charakteristika pro případ, že $f_k \to \infty$ tj. pro $R_v = 0$ (horní křivka).

Pro výpočet časové odezvy reproduktoru použijeme operátorový počet [2]. Pro umožnění výpočtu bylo nutno analogické schéma měniče poněkud zjednodušit.

a) Pro frekvenci budícího signálu v okolí rezonance měniče zanedbáme R_v. Obdobou rovnice (16) pro obraz přenosu měniče v tom případě dostaneme v operátorové formě

$$K_{sa} = Bl \frac{sC_v}{1 + \dfrac{R_e}{R_1} + \dfrac{R_e}{sL_1} + sC_1 R_e + sC_v R_e}. \quad (17)$$

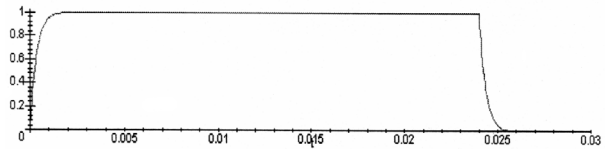
b) Pro budící frekvence značně vyšší neľ je f_k zanedbáme C_v a pak platí

$$K_{sb} = \frac{\dfrac{Bl}{R_v}}{1 + \dfrac{R_e}{R_1} + \dfrac{R_e}{R_v} + \dfrac{R_e}{sL_1} + sC_1 R_e}. \quad (18)$$

Pro budící signál do měniče použijeme časové okno určené výrazem

$$y_o = 1 - e^{-k_x t} + \mathrm{Heaviside}\,(t - T) \left(e^{-k_x (t-T)} - 1 \right). \quad (19)$$

Bylo zvoleno $k_x = 3000$ a délka okna $T = 24\,\mathrm{ms}$. Jeho průběh je na obrázku 4.



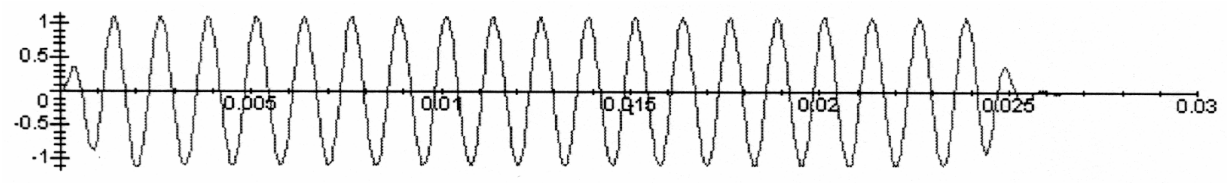
Obrázek 4: Průběh časového okna

Pro budící signál zvolíme výraz

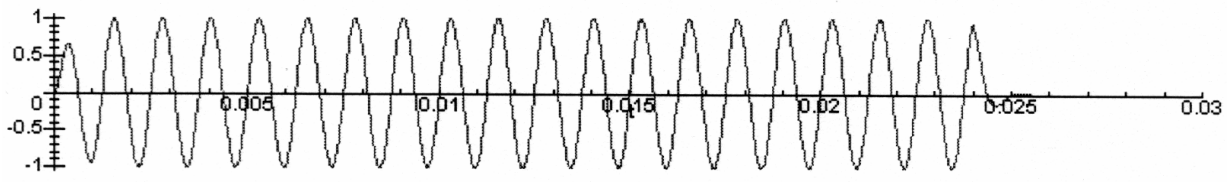
$$y_a = y_0 \sin(2\pi f_a) \quad (20)$$

jehož frekvence $f_a = 800\,\mathrm{Hz}$ leží blízko rezonanční frekvence soustavy a dále

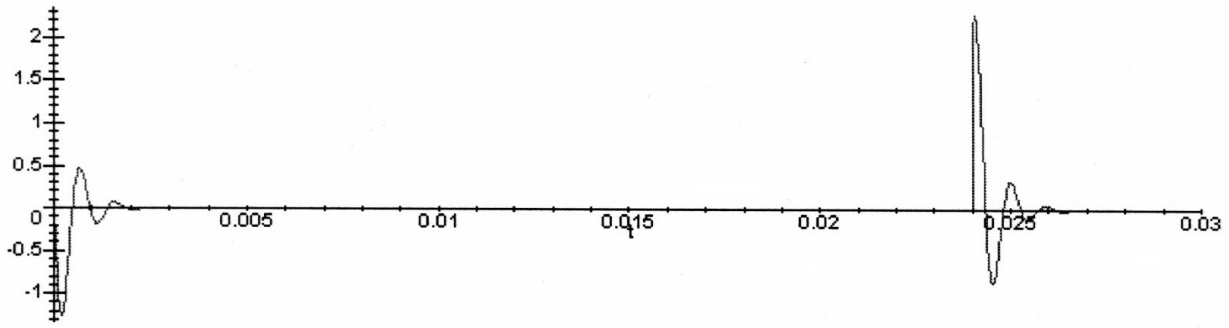
$$y_b = y_0 \sin(2\pi f_b) \quad (21)$$



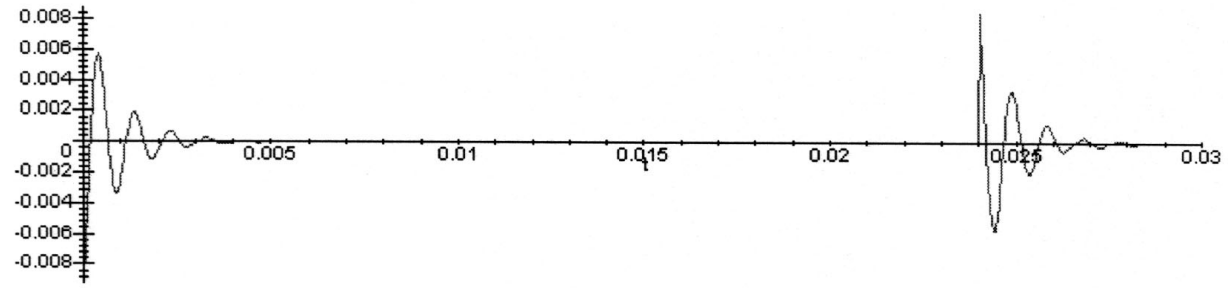
Obrázek 5: Celková výstupní odezva elektrodynamického reproduktoru při budící frekvenci 800 Hz



Obrázek 6: Průběh budícího signálu frekvence 800 Hz



Obrázek 7: Transientní část odezvy elektrodynamického reproduktoru při budicí frekvenci 800 Hz



Obrázek 8: Transientní část výstupní odezvy elektrodynamického reproduktoru při budící frekvenci 5000 Hz

při frekvenci $f_b = 5000\,\mathrm{Hz}$, tedy nad její kritickou frekvencí.

K předmětům y_a nebo y_b najdeme Laplaceovou transformací obrazy

$$Y_{a,b} = \int_0^\infty y_{a,b} e^{-st} dt \tag{22}$$

Součiny $Y_{a,b}K_{sa,b}$ představují L-obrazy výsledné časové odezvy, ke kterým najdeme zpětnou Laplaceovou transformací předměty K_{1a} respektive K_{1b}, vyjadřující průběh výstupního signálu elektrodynamického reproduktoru v časové doméně. Na obrázku 5 je vynesen časový průběh celkové hodnoty K_{1a}. Na obrázku 6 je pro srovnání vynesena křivka, která ukazuje průběh odpovídajícího příslušného budícího signálu. Z porovnání obou průběhů je vidět, že při přenosu došlo k transientnímu zkreslení. Při tom K_{1a} i K_{1b} jsou popsány poměrně složitým výrazem lineárně sdružených funkcí různých argumentů. Když z výrazu pro K_{1a} vybereme komponenty, které nejsou funkcí budicí frekvence f_a, získáme transientní část odezvy. Je vyjádřena goniometrickými funkcemi s frekvencí 1060 Hz, které exponenciálně doznívají. Protože tyto složky v budicím signálu nebyly, představují transientní zkreslení vzniklé při průchodu signálu reproduktorem. Jejich časový průběh je na obrázku 7. Je zřejmé, že tímto způsobem je transientní zkreslení vyjádřeno podstatně zřetelněji, než celkovou odezvou zjištěnou pouhou zpětnou Laplaceovou transformací. Z tohoto důvodu se budeme v dalších úvahách zabývat pouze těmito transientními komponentami.

Obdobným způsobem bylo analyticky zpracováno transientní zkreslení elektrodynamického reproduktoru pro budicí signál o frekvenci 5000 Hz. Na obrázku 8 pak vidíme transientní část odezvy při této frekvenci.

3. Elektrostatický reproduktor

Elektrostatický reproduktor vyvinutý v rámci shora uvedeného grantu použitý pro analytické zjištění transientního zkreslení má obdélníkovou membránu s rozměry $a = 0{,}17$, $b = 0{,}26$. Pro plošnou hmotnost membrány byly zvoleny alternativně hodnoty a) $\rho_{ma} = 0,0034$, nebo b) $\rho_{mb} = 0,005$. Z toho je hmotnost membrány $m_{1a} = 1,503.10^{-4}$, nebo $m_{1b} = 2,21.10^{-4}$. Pro rezonanční frekvenci membrány soustavy $f_f = 5000\,\mathrm{Hz}$ byla určena její poddajnost z výrazu $c_{1a,b} = 1/\omega_r^2 m_{1a,b}$. (V této hodnotě je zahrnuta též poddajnost dutiny za membránou.) Kritická frekvence [1] je dána vzorcem $f_k = c_0/2\pi R$, kde místo R se dosadí hodnota určená rovnicí $\pi R^2 = ab$. Z toho pro rychlost $c_0 = 344\,\mathrm{ms}^{-1}$ plyne $f_k = c_0/2\sqrt{\pi ab} = 462\,\mathrm{Hz}$ a $r_v = 410ab$.

Stanovení mechanického odporu r_1 na membráně [3]: Činitel děrování je $A = 0,325$ a poloměr děr $R_0 = 0,0015$. Počet otvorů v protielektrodě je $N = Aab/\pi R_0^2 = 2030$. Poloměr plochy na odpovídající jednomu otvoru:

$$X_0 = \sqrt{\frac{ab}{N\pi}}. \tag{23}$$

Pro r_1 platí [3]:

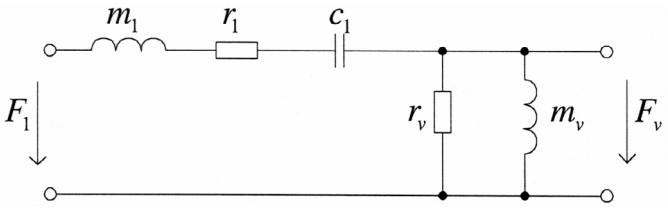
$$r_1 = N\frac{6\mu\pi X_0^4}{h^3}\beta, \tag{24}$$

kde

$$\beta = \ln\frac{X_0}{R_0} - \frac{3}{4} + \frac{R_0^2}{X_0^2} - \frac{R_0^4}{4X_0^4}, \tag{25}$$

kde $\mu = 1,82\,10^{-5}$ značí dynamickou viskozitu vzduchu a $h = 1,2\,10^{-4}$ vzdálenost membrány od protielektrody.

Analogické schéma mechanické části soustavy je na obrázku 9.



Obrázek 9: Analogické schéma mechanické části elektrostatického reproduktoru

Protože síla $F_1 = k_b u_1$, tedy je úměrná budícímu napětí u_1 viz [1], můžeme pro přenos psát

$$K_{2f} = \frac{F_v(\omega)}{F_1(\omega)} = \frac{Z_v(\omega)}{\mathrm{j}\omega m_1 + r_1 + \dfrac{1}{\mathrm{j}\omega c_1} + Z_v(\omega)}, \tag{26}$$

kde Z_v značí opět vyzařovací impedanci danou vztahem (3).

Vypočtené průběhy frekvenční charakteristiky elektrostatického reproduktoru jsou pro obě zvolené plošné hmotnosti membrány na obrázku 10.

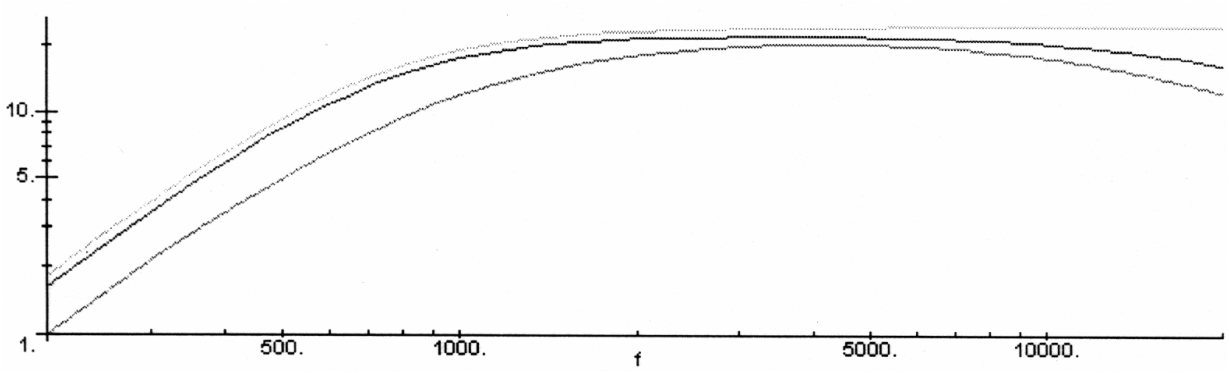
Pro výpočet transientních jevů byl opět použit operátorový počet. Pro umožnění výpočtu bylo nutno i zde zavést určité zjednodušení. Lze zřejmě zanedbat, jak lze soudit z průběhu frekvenčních charakteristik v obrázku 10, hodnotu m_{1a}, resp. m_{1b} . V operátorové formě pak pro přenos soustavy získáme z rov. (26) vztah:

$$K_s = \frac{F_v(s)}{F_1(s)} = \frac{Z_v(s)}{r_1 + \dfrac{1}{sc_1} + Z_v(s)}, \tag{27}$$

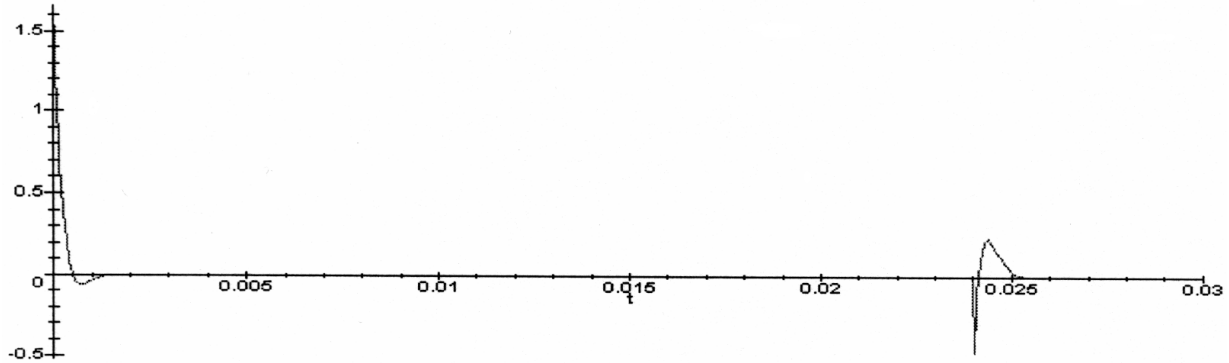
kde

$$\frac{1}{Z_v(s)} = \frac{1}{r_v} + \frac{1}{sm_v}. \tag{28}$$

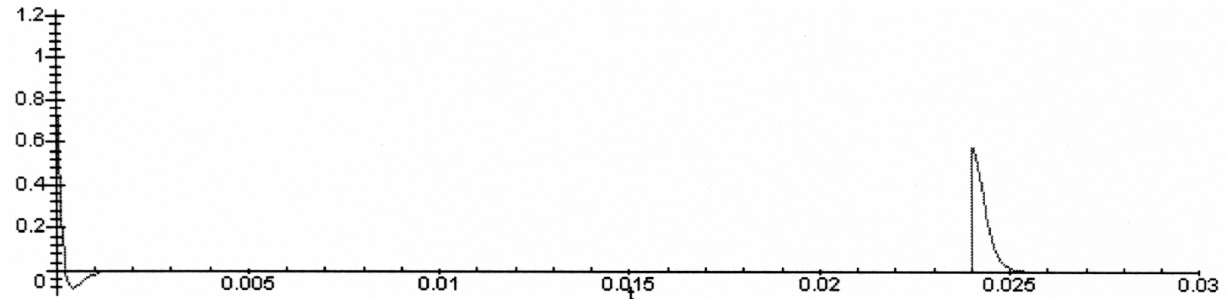
Protože používáme i zde pro srovnání tytéž budicí signály jako pro reproduktor elektrodynamický, platí pro ně i v tomto případě výrazy y_a a y_b dané rovnicemi (20) a (21) a jejich obrazy Y_a respektive Y_b podle rovnice (22). Součiny $Y_{a,b}K_s$ potom jsou L-obrazy výsledné časové odezvy, navrženého elektrostatického reproduktoru. K těmto obrazům najdeme zpětnou Laplaceovou transformací předměty K_{2a} respektive K_{2b}, vyjadřující výstupní signál v časové doméně. Z těchto výrazů vybereme opět pouze transientní část, a to jak pro membránu s plošnou hmotností 0,0034, tak také 0,005.



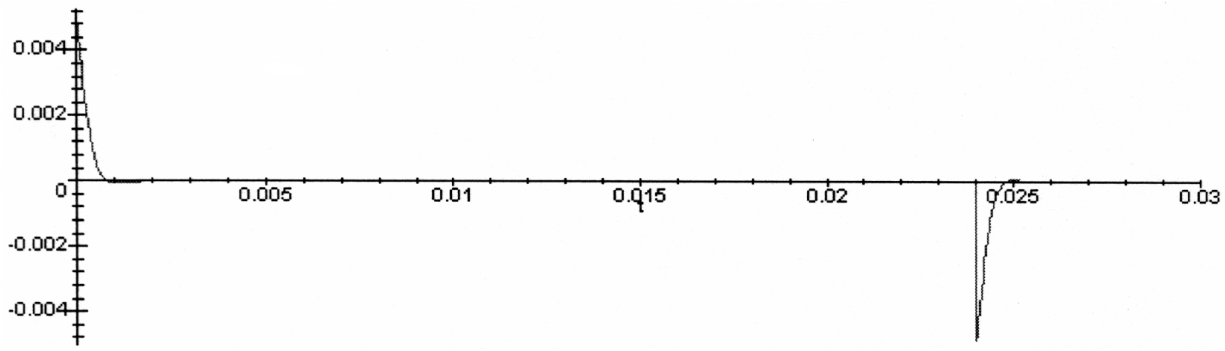
Obrázek 10: Frekvenční charakteristika elektrostatického reproduktoru. Střední křivka platí pro plošnou hmotnost membrány 0,0034, dolní pro 0,005. Horní křivka vychází při zanedbání m_1



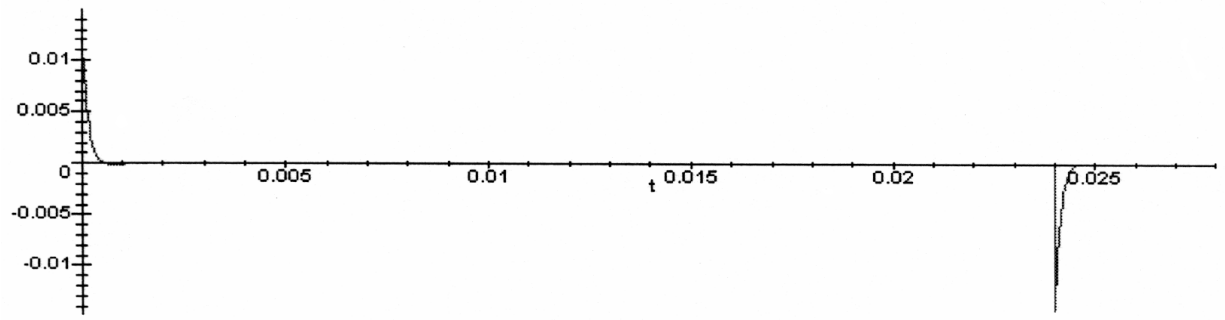
Obrázek 11: Transientní část odezvy elektrostatického reproduktoru při budící frekvenci 800 Hz, pro plošnou hmotnost membrány 0,0034



Obrázek 12: Transientní část odezvy elektrostatického reproduktoru při budící frekvenci 800 Hz, pro plošnou hmotnost membrány 0,005



Obrázek 13: Transientní část odezvy elektrostatického reproduktoru při budící frekvenci 5000 Hz. Plošná hmotnost membrány 0,0034



Obrázek 14: Transientní část odezvy elektrostatického reproduktoru při budící frekvenci 5000 Hz. Plošná hmotnost membrány 0,005

To provedeme pro obě zvolené frekvence budícího signálu tedy 800 Hz a 5000 Hz. Výsledky jsou vyneseny na následujících obrázcích.

Na obrázku 11 je znázorněna transientní část časové odezvy na signál o frekvenci 800 Hz, pro membránu s plošnou hmotností 0,0034. Na obrázku 12 je vynesena transientní část odezvy při téže budicí frekvenci, ale pro membránu s plošnou hmotností 0,005. Na obrázcích 13 a 14 jsou obdobně znázorněny transientní odezvy, ale pro budicí frekvenci 5000 Hz. Všechny zde uvedené křivky transientních částí odezvy platí pro hodnotu stacionární části výstupního signálu jedna.

4. Závěr

Uvedené výsledky potvrzují skutečnost uvedenou v úvodu, že elektrostatické měniče jsou co do transientního skreslení signálu mnohem výhodnější, než běžně používané reproduktory elektrodynamické. Za povšimnutí stojí při tom zejména to, že frekvence transientních jevů, které se vyskytují při náběhu a doběhu signálu v podstatě nezávisí na frekvenci přenášeného signálu, ale jsou dány vlastnostmi měniče.

Poděkování

Chtěl bych rád poděkovat profesorovi Zdeňku Škvorovi za pečlivé prostudování této práce a cenná doporučení pro její publikaci. Rozbor byl proveden v rámci prací na grantovém projektu číslo 102/00/1661 grantové agentury České republiky.

Reference

[1] Merhaut, J.: *Teoretické základy elektroakustiky*, Academia Praha 1985.

[2] Angot, A.: *Užitá matematika pro elektrotechnické inženýry*, SNTL, Praha 1960.

[3] Škvor, Z.: *Vibrating Systems and their Equivalent Circuits*, Academia Praha 1991.

Vícevstupové metody redukce šumů v řeči

Pham Quang Hung a Pavel Sovka

ČVUT-FEL, Technická 2, 16627 Praha 6
email: [hung, sovka]@feld.cvut.cz

The paper is devoted to the brief description of noise reduction systems used for noise reduction in signals contaminated by additive noises. Furthermore the new family of coherence-based noise reduction systems is described. Comparison and performance limits of chosen methods are given.

1. Úvod

Článek je věnován stručnému přehledu stavu problematiky potlačování aditivních šumů přítomných v akustických signálech snímaných mikrofonem a podrobněji popisuje jednu rodinu metod, které se zdají být efektivní pro úspěšnou redukci šumů v řeči.

Problematika redukce šumů v akustických, především řečových signálech je intenzivně řešena již po čtyři desetiletí. Těsně souvisí s rozvojem technologií zpracování, přenosu a rozpoznávání řeči. Na rozdíl od metod aktivního potlačování hluků [1], [2] používaných v pomůckách ochrany sluchu, tlumení hluku způsobeného vibracemi (např. v jedoucím automobilu) nebo vytváření oblastí ticha, není třeba pro potlačení hluku generovat akustický „antihluk“, neboť potlačování není prováděno v akustické oblasti, ale v „elektrické“. To znamená, že šum je redukován přímo v signálu získaného na výstupu mikrofonu. Způsoby redukce šumu jsou dva. Buď je šum redukován filtrací (vybraná frekvenční pásma jsou tlumena), nebo kompenzací, kdy je vytvořen „antišum“ a ten je přičítán ke směsi signálu se šumem. Metody redukce hluku jsou již od počátku jejich vývoje realizovány výhradně číslicovými prostředky zpracování signálu. To znamená, že směs užitečného signálu s aditivním šumem je digitalizována a následně zpracována procesorem. Podle aplikace je zvýrazněný signál s redukovaným šumem buď (po případném kódování [3] a přenosu) opětovně rekonstruován nebo dále zpracováván. První případ je typický pro systémy redukce hluků v signálu, které mají na svém vstupu i výstupu signál (např. pro prostředky mobilní telefonie), druhý případ lze nalézt např. v systémech rozpoznávání řeči [4].

Oba přístupy potlačování hluku spolu těsně souvisejí, používají podobných algoritmů, proto i terminologie je v některých případech podobná. Např. zkratka ANC má dvojí význam: aktivní potlačovánů hluků (Active Noise Cancellation) nebo označuje adaptivní systém pro kompenzaci hluků (Adaptive Noise Cancellation). Druhý význam je často používán právě pro zmíněné systémy „elektrického“ a nikoliv akustického potlačování šumů. Tyto metody nepotřebují měniče pro generování „antihluku“, lze je proto simulovat snadněji než metody aktivního potlačování.

2. Rozdělení metod redukce šumů

V celém dalším textu budou popisovány výhradně systémy potlačování šumů použité pro redukci šumů v řeči snímané mikrofony.

Podle použitého uspořádání lze metody redukce šumů rozdělit do tří základních skupin:

- metody s jedním mikrofonem, označované jako jednovstupové,
- metody se dvěma mikrofony, označované jako dvouvstupové,
- metody s více než dvěma mikrofony, označované jako vícevstupové nebo systémy s mikrofonním polem.

Lze samozřejmě namítnout, že metody vícevstupové jsou zobecněním dvouvstupových, a proto nemá smysl dvouvstupové uvádět samostatně. Skutečností je, že dvouvstupové metody mají právo na samostatnou skupinu alespoň ze dvou důvodů. Jedním je, že použitý fyzikální princip redukce šumů dvouvstupových a vícevstupových se v realizovaných systémech liší (z důvodu účinnosti) a rovněž realizační nároky obou skupin se značně rozcházejí.

Typické vlastnosti uvedených systémů lze zhruba shrnout následující způsobem:

- jednovstupové systémy jsou jednoduché, proto levné a dobře propracované. Jsou v aplikacích (redukce šumů pro přenosové systémy, rozpoznávání řeči) často používány. Zvýšení odstupu signálu od šumu (SNRE-Signal to Noise Ratio Enhancement) bývá v rozsahu 4-8 dB, v závislosti na odstupu zpracovávaného signálu od šumu (SNR-Signal to Noise Ratio). Nežádoucím důsledkem redukce šumu je vznik zkreslení řeči a vznik reziduálních šumů, které jsou pro svůj charakter nazývány hudebními tóny (šumy).
- dvouvstupové systémy jsou stále ještě relativně jednoduché. Redukce šumu je podobná jako u jednovstupových systémů, ale v důsledku lepšího využití informace v signálu mohou poskytovat menší zkreslení řeči i úroveň reziduálních šumů. Vzhledem k možnosti lokalizovat směr přicházejícího signálu a relativní jednoduchost jsou používány v pomůckách pro sluchově postižené.

- vícevstupové systémy umožňují dosažení nejlepších (neznamená vždy vyhovujících) parametrů ze všech tří skupin. Omezení jejich účinnosti je dáno počtem a geometrií pole mikrofonů. V praxi se (z ekonomických důvodů) používají nejčastěji čtyři mikrofony.

Optimalizace libovolného systému je vždy dána kompromisem mezi požadovanou úrovní redukce šumů, zkreslením řeči a úrovní reziduálních šumů. Rovněž vybraná aplikace a typ zpracovávaného signálu (např. řeč/hudba) určují váhu jednotlivých požadavků na optimalizaci parametrů systému. Například nároky na zkreslení řeči pro účely přenosu řeči a rozpoznávání řeči se značně liší. Důsledkem této rozmanitosti je skutečnost, že neexistuje univerzální metoda ani postup její optimalizace a dokonce ani postup jejího použití pro zpracování signálů s aditivním šumem.

3. Základní principy redukce šumů v akustických signálech

Nyní stručně a bez nároků na přesnost a úplnost budou uvedeny základní principy nejčastěji používané v uvedených třech skupinách metod. Výběr principů je podřízen potřebám tohoto článku.

Obecně lze tvrdit, že redukce šumu obsaženého v signálu vyžaduje rozlišení šumu od signálu na základě jejich odlišných charakteristik nebo odlišného časového vývoje charakteristik popřípadě odlišností signálů ve směru šíření a časovém zpoždění. Nejčastěji používanými charakteristikami jsou korelace, spektra a koherence.

- Pro jednovstupové metody byla původně použita autokorelace, později spektrum popř. spektrální hustota. Princip jedné z prvních a často používaných metod označované jako spektrální odečítání (např. [5], [6]) spočívá v odhadu spektrální hustoty šumu v pauzách řeči a její následném odečtení od spektra vstupního signálu. Jiný přístup je založen na rozkladu signálu do nepřekrývajících se frekvenčních pásem a tlumením těch pásem, která jsou zasažena silným šumem [7]. Oba přístupy upravují spektrum vstupního signálu tak, aby byly zvýrazněny jeho části nesoucí informaci o signálu a potlačeny části maskované šumem. V některých případech jsou oba přístupy identické. Pro rozklad do pásem se používá banka filtrů s lineárním i nelineárním frekvenčním krokem, diskrétní Fourierova transformace, nověji též vlnková transformace. Pro rekonstrukci signálu s redukovaným šumem se požívá fáze vstupního signálu. Další zlepšení metody spektrálního odečítání a použití nelineárních funkcí pro zvýšení kvality vedly k metodám, které se dnes považují za etalon jednovstupových metod [8], [9]. Podle autora se nazývají Ephraimovy metody. Další modifikace vedly k efektivní, často používané realizaci metody redukce šumů [10], která bývá označena jako metoda Akbariho-Aziraniho.

  Shrňme problematické vlastnosti spektrálního odečítání (podobné vlastnosti vykazuje i metoda tlumení frekvenčních pásem).

  – Je zřejmé, že v případě nepřesného odhadu šumového pozadí dochází při jeho odečtení ke vzniku chyb - reziduálních šumů. Podobná situace nastane, změní-li se parametry šumového pozadí během přítomnosti řeči, kdy nelze odhad šumu aktualizovat. Pro signály s rychlými změnami šumového pozadí se proto účinnost této metody snižuje.
  – Uvedený postup vyžaduje spolehlivý detektor řeči a pauzy (VAD-Voice Activity Detector) [11], [12]. To je v případě nízkého odstupu signálu od šumu téměř neřešitelný problém. Především přesná detekce počátků řeči je při přítomnosti šumu nemožná. Nezachycení řečové aktivity detektorem způsobí odečtení průměrného spektra řeči od spektra signálu a ve výsledku zkreslení řeči až ozvěnu.

  Poznámka: Pro rychlé změny šumového pozadí, popřípadě pro signály bez pauz lze použít jednovstupové metody využívající jiný způsob měření šumového pozadí. Odhad spektra šumu lze získat z minim spektra získaných z delšího úseku signálu [13], [14] nebo pomocí filtrace přizpůsobeným filtrem [15] . Ovšem i tyto metody produkují reziduální šumy a zkreslení řeči.

- Dvouvstupové metody, ze kterých vychází tato práce využívají informaci o vzájemných vazbách mezi dvěma signály, tedy vzájemnou korelaci, vzájemné spektrum a koherenci. V důsledku použití dvou vstupů je možné využít rovněž prostorovou informaci. Právě dvouvstupové metody velmi často využívají normované vzájemné spektrum, tedy koherenci (přesněji kvadrát jejího modulu: MSC-Magnitude Squared Coherence). Mezní hodnoty této funkce jsou 0 pro dva nekorelované (nekoherentní) šumy a 1 pro dva plně korelované (koherentní) řečové signály.

  V realizacích jsou používány často (ale nikoliv výlučně) dva přístupy.

  – První přístup je založen na redukci šumu přizpůsobeným filtrem, jehož parametry jsou určeny právě mírou vzájemné závislosti obou signálů např. [16].
  – Druhý způsob využívá vzájemného spektra nebo koherence pro získání spektra šumového pozadí [17], [18], které je následně odečteno od spektra vstupního signálu.

  Oba přístupy obecně nevyžadují explicitní detekci řeči, i když v některých modifikacích je použita. Tyto přístupy vyžadují, aby aditivní šumy byly difuzní

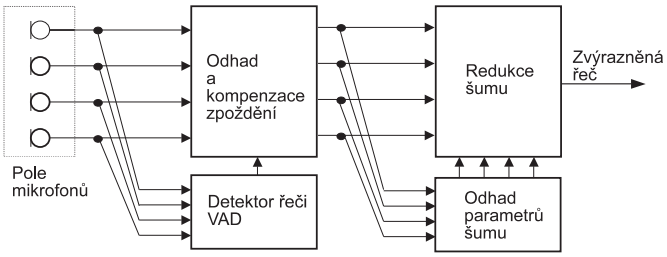
(metoda [16] a jedna modifikace z druhé skupiny [17]). Existuje ale i modifikace, která za cenu požadavku detekce řeči nabízí redukci koherentních šumů [18], což je například typický případ některých hluků v jedoucím automobilu. Právě tento případ je zajímavý pro použití telefonu v automobilu jakož i ovládání některých funkcí automobilu.

- Vícevstupové systémy využívají převážně prostorovou informaci.
  - Jeden z postupů redukuje šum tím, že se sčítají spektra synchronizovaných vstupních signálů, čímž výkon koherentních složek signálu roste rychleji než výkon nekoherentních složek. V důsledku toho dochází ke zvětšení odstupu signálu od šumu. Tento přístup vyžaduje odhad a kompenzaci zpoždění mezi jednotlivými signály[1] [19] (viz obr. 1). Pro tento odhad bývají používány korelace, koherence nebo statistiky vyšších řádů. Opět je nutné použít detektor řeči. Tyto systémy bývají doplněny filtrací pro redukci koherentních složek šumu a v literatuře jsou označovány jako GSC-Generalized Sidelobe-Canceller např. [20].
  - Jiný možný přístup (použitý v našem případě) je zobecnění dvouvstupové struktury, která odhad spektra šumového pozadí získaný pomocí vzájemného spektra nebo koherence používá pro potlačení složek signálu maskovaných šumem (viz první tři bloky na obr. 2).

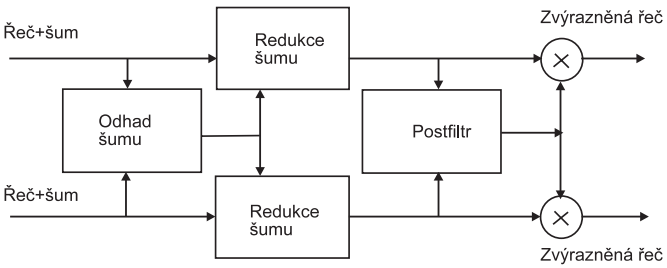
Poznámka: Přístup používaný v případech, kdy lze získat odhad šumu přítomného signálu pomocí referečního čidla a pak jej odečíst od vstupního signálu v časové (nebo spektrální) oblasti je označován jako kompenzační metoda. Jedná se o již zmíněné ANC navržené Widrowem [21]. Tento systém využívá dvě a více čidel a nepožaduje detektor přítomnosti signálu. Jeden ze vstupů, označený jako primární by měl obsahovat signál se šumem, ostatní vstupy (referenční) by měly obsahovat pouze šumy. Pro správnou funkci systém vyžaduje korelovanost šumů ve všech vstupech a zároveň jejich nekorelovanost se signálem. Právě ANC provádí dodatečnou redukci koherentních šumů ve vícevstupových metodách. Zatímco součet spekter vstupních signálů představuje primární vstup ANC, jejich rozdíl formuje referenci šumu.

Existuje řada dalších metod. Namátkou jmenujme iterační metody [7], [22], které původní odhad parametrů šumu iterativně upřesňují a metody tvarování spektra [2] použitelné v pomůckách pro sluchově postižené. Další používané přístupy redukce šumů v řeči lze nalézt např. v knize [23].

[1] Kompenzace zpoždění signálů způsobuje, že pole mikrofonů vykazuje maximální citlivost ve směru dopadu dominantní koherentní složky signálu a nižší citlivost v ostatních směrech. Tím se při redukci nekoherentních šumů uplatňuje směrová charakteristika pole mikrofonů podobně jako u anténních soustav.



Obrázek 1: Princip vícevstupových metod s kompenzací zpoždění



Obrázek 2: Dvouvstupový systém redukce šumu

4. Popis návrhu dvouvstupové metody

Za základ vývoje systémů redukce šumů byla vybrána struktura navržená v [17] (Dorbeckerova struktura), která je zobrazena na obr. 2. Ve struktuře lze rozpoznat tři základní bloky:

- první blok, označený jako *Odhad šumu*, pomocí vzájemného spektra nebo koherence odhaduje spektrum šumového pozadí,
- další dva bloky, označené jako *Redukce šumu*, provádějí úpravu spekter vstupních signálů pomocí odhadnutého spektra šumu a spektra vstupních signálů,
- poslední blok, označený jako *Postfiltr*, poskytuje dodatečné vyhlazení spekter signálů, čímž snižuje úroveň reziduálních šumů. Není-li potřeba použít oba výstupy odděleně, lze součtem spekter obou větví dále zvýšit odstup užitečného signálu od nekoherentních šumů.

Celý systém je realizován ve frekvenční oblasti. Vstupní signál je segmentován na úseky délky N vzorků s překrytím L vzorků (typicky $N = 256$, $L = 128$ pro vzorkovací frekvenci 10-12 kHz). Segmenty jsou pomocí diskrétní Fourierovy transformace převedeny do frekvenční oblasti. Signál je rekonstruován ze spektra pomocí inverzní diskrétní Fourierovy transformace a metodou sčítání přesahů (např. [4], [23]). Tyto bloky nejsou v obr. 2 z důvodu přehlednosti vyznačeny.

4.1. Optimalizace dílčích bloků systému

Původní postup odhadu šumu [17] předpokládal difúzní nebo nekoherentní šumy, což je v některých aplikacích ne-

přijatelné omezení, proto byl blok odhadu šumu nahrazen postupem uvedeným v [18] (Simmerova metoda). Tento postup odhaduje parametry šumu pomocí koherence a připouští i koherentní typy šumů.

Odhad spektrální hustoty $P_{NN}(f)$ šumu na výstupu bloku *Odhad šumu* je dán vztahem

$$P_{NN}(f) = H(f)\, P_R(f), \tag{1}$$

kde $P_R(f)$ je spektrální hustota rozdílu rozdílu spekter $X_1(f)$ a $X_2(f)$ obou vstupních signálů.

Přenosová funkce $H(f)$ je určena

$$H\,(f) = \frac{1 + \Re\{C_{X_1X_2}\,(f)\}}{1 - \Re\{C_{X_1X_2}\,(f)\}}, \tag{2}$$

kde symbol $\Re$ představuje reálnou část koherence $C_{X_1X_2}$ definované spektrálními hustotami $P_{X_iX_j}$ vstupních signálů

$$C_{X_1X_2}(f) = \frac{P_{X_1X_2}(f)}{\sqrt{P_{X_1X_1}(f)P_{X_2X_2}(f)}}. \tag{3}$$

Spektrální hustoty jsou v tomto případě odhadovány pomocí rekuretního vztahu prvého řádu s vhodně zvolenou časovou konstantou, jejíž velikost závisí na rychlosti změn parametrů signálů.

Dalším parametrem pro optimalizaci systému je blok provádějící redukci šumu.

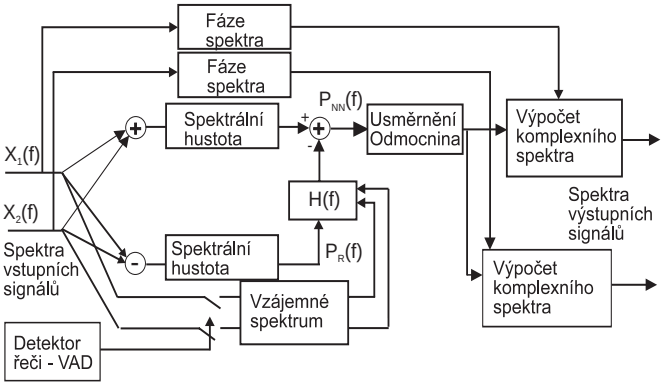
V původní práci [17] byl použit Ephraimův algoritmus [9]. Z důvodů úspory operací i zvýšení redukce šumu byly testovány i další algoritmy. Jako kompromis mezi výpočetními nároky a vlastnostmi celého systému lze pro realizaci tohoto bloku použít zjednodušenou verzi spektrálního odečítání. Pro relativně nízké výpočetní nároky je v současné době testována pro použití v pomůckách pro sluchově postižené. Nejlepších výsledků lze dosáhnout s metodou Akbariho-Aziraniho [10].

Posledním krokem je posfiltrace obou signálů prováděná pomocí adaptivní Wienerovy filtrace podle vztahu [17]

$$W(f) = \frac{\left|\hat{P}_{\tilde{X}_1\tilde{X}_2}(f)\right|}{\hat{P}_{\tilde{X}_1\tilde{X}_1}(f)} \quad , \tag{4}$$

kde $\tilde{X}_1$ a $\tilde{X}_2$ jsou spektra signálů na výstupu bloků *Redukce šumu*.

Výsledná struktura dvou základních bloků systému, tj. odhad šumu a jeho redukce ve vstupních signálech pomocí spektrálního odečítání je zobrazena na obr. 3. Pro přehlednost byl vynechán blok provádějící postfiltraci, který by byl zařazen na výstup soustavy. Je-li místo spektrálního odečítání použita jiná metoda, jsou symbol rozdílu a blok odmocniny nahrazeny blokem provádějící násobení spektrálních čar koeficienty odvozenými z odhadu šumu (viz blok *Redukce šumu* na obr. 2).



Obrázek 3: Navržený dvouvstupový systém redukce šumu

5. Zobecnění dvouvstupové metody

Dvouvstupová struktura uvedená v předchozí sekci neposkytuje vždy dostatečnou redukci šumu. Proto bylo provedeno zobecnění této struktury na čtyřvstupovou. Zobecnění spočívá ve vytvoření všech možných dvojic signálů z příslušných vstupů, které jsou zpracovány podle obr. 2. Nárůst počtu operací je v důsledku využívání mezivýsledků zhruba poloviční, než by odpovídalo počtu nárůstu danému počtem všech dvojic. Nicméně počet operací lze dále redukovat vynecháním těch dvojic signálů, které nepředstavují významný příspěvek pro redukci šumu. Efektivní algoritmus výběru dvojic signálů je v současné době vyvíjen. Experimenty ukázaly, že pro vyrovnané dynamické poměry na vstupech soustavy lze použít i pouhý náhodný výběr dvojic signálů. Tento postup umožňuje zachovat vyšší redukci šumu než které dosahuje dvouvstupová struktura při nepatrně větším objemu operací. Více podrobností a detailní rozbor výpočetních nároků lze nalézt v práci [24].

6. Experimenty a výsledky

Dvouvstupový a čtyřvstupový systém popsaný v předchozí sekci byl testován pomocí reálných šumů a řeči nahraných v jedoucím automobilu. Pro kvantitativní vyhodnocení bylo ovšem třeba použít řeč $s[n]$ sejmutou mikrofonem ve stojícím automobilu a sečíst ji se šumem $n[n]$ jedoucího automobilu[2]. Výstupem systému provádějícího redukci šumu je signál $\hat{s}[n]$ (horní blok na obr. 4).

Byly hodnoceny

- průměrná (segmentální) hodnota zvýšení odstupu signálu od šumu (SSNRE-Segmental Signal to Noise Ratio Enhancement)

$$SSNRE = \frac{1}{M}\sum_{i=1}^{M} SNRE(i), \tag{5}$$

kde M je počet segmentů vstupního signálu, ve kterých je přítomna řeč a $SNRE(i)$ je hodnota SNRE

[2]Podmínky nahrávání signálů, jejich parametry jakož i použité konfigurace mikrofonů lze opět nalézt v [24].

příslušná i-tému segmentu, která je pro libovolný segment dána vztahem

$$SNRE = 10\log\frac{\sum_{k=1}^{N} n^2\,[k\,]}{\sum_{k=1}^{N}\left(s\,[k\,]-\hat{s}\,[k\,]\right)^2}. \qquad (6)$$

N je počet vzorků segmentu (typicky $N = 256$).

- zkreslení řeči (SD-Speech Distortion)

$$SD = 10\log\frac{\sum_{k=1}^{N}(\tilde{s}[k]-s[k])^2}{\sum_{k=1}^{N} s^2[k]}, \qquad (7)$$

kde $\tilde{s}[k]$ je výstup systému, do kterého vstupuje pouze čistá řeč a jehož parametry jsou dány systémem provádějícím zpracování vstupního signálu (prostřední blok na obr. 4). Pro hodnocení je použita průměrná hodnota SegSD určená obdobně jako SSNRE (5) z jednotlivých segmentů signálu.

- potlačení šumu (NR-Noise Reduction) bez ohledu na přítomnost řeči

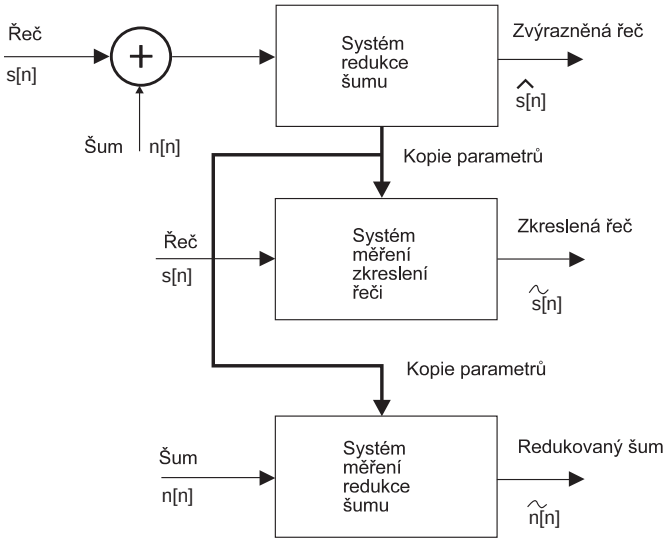
$$NR = 10\log\frac{\sum_{k=1}^{N} n^2[k]}{\sum_{k=1}^{N}\tilde{n}^2[k]}, \qquad (8)$$

kde $\tilde{n}[k]$ je výstup systému, do kterého vstupuje pouze šum a jehož parametry jsou dány systémem provádějícím zpracování vstupního signálu (spodní blok na obr. 4). Opět je použita průměrná hodnota SegNR.

Poznámka: Je nutné zdůraznit, že kritéria SD a NR lze vyčíslit popsaným způsobem pouze pro lineární metody, např. Akbari-Azirani, Ephraim-Malah. Nelze je získat pro spektrální odečítání, které vzhledem k použití absolutní hodnoty není lineární.

Hodnocení metod redukce šumů bylo provedeno pro dva základní případy. V prvním případě byl použit téměř stacionární šum nahraný za jízdy v automobilu, ve druhém případě energie šumu postupně narůstala. Výsledky ukazující zvětšení průměrného odstupu signálu od šumu (SSNRE) pro spektrální odečítání a metodu Akbariho-Aziraniho jsou uvedeny v tab. 1. Výsledky dosažené jinými metodami v podstatě leží mezi hodnotami získanými pomocí těchto dvou metod.

Hodnoty udávající zkreslení řeči (SegSD) jsou bez použití postfiltrace v okolí 2 dB. Zkreslení řeči je v tomto případě slyšitelné. Při použití postfiltrace klesne zkreslení pod hranici 0.5 dB a přestává být slyšitelné. Velikost průměrné redukce šumu je při použití postfiltrace větší než



Obrázek 4: Struktura pro hodnocení zkreslení řeči a potlačení šumu

Dvouvstupová metoda	P0	P1	P1A1
spektrální odečítání	6.5	7	7.3
Akbari-Azirani	8.5	8.7	9.2

Tabulka 1: Výsledky SSNRE [dB] pro dvě vybrané dvouvstupové metody a stacionární šum: P0-bez postfiltrace a sečtení výstupů, P1-s postfiltrací a bez sečtení výstupů, P1A1-s postfiltrací a sečtením výstupů

40 dB, bez ní je přibližně poloviční. Pro čtyřvstupové metody s postfiltrací je hodnota SSNRE přibližně o 2 dB větší než pro dvouvstupové metody, přičemž zkreslení řeči i úroveň reziduálních šumů je nižší.

Z porovnání výsledků je zřejmé, že dodatečné vyhlazení upravených spekter signálů postfiltrací, popřípadě jejich sečtení zlepšuje kvalitu výstupního signálu, to znamená snižuje zkreslení řeči při větší redukci šumu. Větší redukce šumu má rovněž za následek menší úroveň reziduálních tónů. Výsledky pro nestacionární signály byly blízké uvedeným výsledkům. Tendence změn v hodnotách kritérií dobře korelovala s orientačními poslechy prováděnými skupinou posluchačů.

7. Závěr

V textu byla představena rodina metod redukce šumů. Metodám je společné, že pro odhad parametrů šumů využívají koherenci získanou z více vstupních signálů. To umožňuje dosáhnout dostatečné redukce šumů při nízkém zkreslení řeči a nízké úrovni reziduálních šumů. Dosažené výsledky ukazují, že uvedené metody lze použít jak pro účely předzpracování řeči před kódováním a přenosem, tak i pro účely rozpoznávání řeči. Podrobný rozbor výpočetních nároků a implementace dílčích částí ukázaly, že pro realizaci

těchto metod v reálném čase vyhovuje výpočetní i paměťová kapacita jednoho moderního signálového procesoru.

Poděkování

Tato práce byla podporována komplexním grantem „Hlasové technologie v podpoře informační společnosti“, GA ČR 102/02/0124 a výzkumným záměrem „Transdisciplinární výzkum v oblasti biomedicínského inženýrství“, MSM 210000012.

Reference

[1] Kuo, S. M., Morgan, D. R. : *Active Noise Control Systems.* John Wiley & Sons, Inc., New York, 1996.

[2] Nelson, P. A. : *Active Control of Sound.* Academic Press, New York, 1993.

[3] Kondoz, A. M. : *Digital Speech. Coding for low bit rate communications systems.* John Willey & Sons., New York, 1996.

[4] Psutka, J. : *Komunikace s počítačem mluvenou řečí.* Academia Praha, 1995.

[5] Boll, S. F. : Suppression of acoustic noise in speech using spectral subtraction. *ASSP*, ASSP-27(2):113–120, April 1979.

[6] Kang, G. S. Fransen, L. J. : Quality improvement of LPC-processed noisy speech by using spectral subtraction. *ASSP*, ASSP-37(6):939–942, June 1989.

[7] Lim, J. S. : *Speech Enhancemment.* Prentice-Hall, Inc., 1983 Collection of reprints.

[8] Ephraim, Y. Malah, D. : Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *ASSP*, Vol. ASSP-32(6):1109–1121, December 1984.

[9] Ephraim, Y. Malah, D. : Speech enhancement using a minimum mean-square log-spectral amplitude estimator. *ASSP*, Vol. ASSP-33(No. 6):443–445, 1985.

[10] Akbari, A., Bouquin, R. L., Bouquin, G. : Speech enhancement using a wiener filter under signal presence uncertainty. *EUSIPCO'96*, 1996.

[11] Haigh, J. A., Mason, J. S. : A voice activity detector based on cepstral analysis. *EUROSPEECH'93 - Proceedings of the 3rd European Conference on Speech, Communication, and Technology*, str. 1103–1106, Berlin, September 1993.

[12] Sovka, P. Pollák, P. : The study of speech/pause detectors for speech enhancements methods. *Proceedings of the 4th European Conference on Speech Communication and Technology - EUROSPEECH'95*, str. 1575–1578, Madrid, Spain, September 1995.

[13] Martin, R. : Spectral subtraction based on minimum statistics. *Proceedings of EUSIPCO-94 Seventh European Signal Processing Conference*, str. 1182–1185, Edinburgh, Scotland, U.K., September 1994.

[14] Doblinger, G. : Computationally effecient speech enhancement by spectral minima tracking in subbands. *EUROSPEECH'95 - Proceedings of the 4th European Conference on Speech Technology and Communication*, str. 1513, Madrid, Spain, September 1995.

[15] Sovka, P., Pollák, P., Kybic, J. : Extended spectral subtraction. *EUSIPCO'96*, Trieste, September 1996.

[16] Bouquin, R. L., Faucon, G. : Using the coherence function for noise reduction. IEEE Proceedings-139(3):276–280, June 1992.

[17] Dorbecker, M., Ernst, S. : Combination of two-channel spectral subtarction and adaptive Wiener post-filtering for noise reduction and deverberation. *IWAENC'97 - Proceedings of the 2nd International Workshop on Echo and Noise Cancellation*, 1997.

[18] Meyer, J., Simmer, K. U., Kammeyer, K. D. : Comparison of one- and two-channel noise estimation techniques. *IWAENC'97 - Proceedings of the 2nd International Workshop on Echo and Noise Cancellation*, 1997.

[19] Simmer, K. U., Kuczynski, P., Wasiljeff, A. : Time delay compensation for adaptive multichannel speech enhancement systems. *Proceedings of URSI 92*, Paris, 1992.

[20] Bitzer, J., Simmer, K. U., Kammeyer, K. D. : Multichannel noise reductionalgorithms and theoretical limits. *EUSIPCO'98*, str. 105–108, Greece, Sep. 1998.

[21] Widrow, B., Stearns, S. D. : *Adaptive signal processing.* Prentice-Hall, Inc., New Jersey, 1985.

[22] Proakis, J. G., Rader, C. M., Ling, F., Nikias, C. L. : *Advanced Digital Signal Processing.* Macmillan Publishing Company, New York, 1992.

[23] Vaseghi, S. V. : *Advanced Signal Processing and Digital Noise Reduction.* Wiley Teubner, New York, 1995.

[24] Hung, P. : *Optimization and Implementation of Multi-channel Speech Enhancement Methods.* PhD thesis, FEE-CTU in Prague, 2002.

Kontextově závislé modely fonémů

Jan Novotný

FEL ČVUT, katedra teorie obvodů
Technická 2, 166 27 Praha 6
e-mail: novotnj2@feld.cvut.cz

A progressive way of Hidden Markov Models training lately used in speech processing is described in this paper. The description is focused on the training of monophone and triphone HMMs considering their possible application advantages. This way of HMMs training does not require use of speech database with manually labeled time marks. The database just with word level label files is needed for this purpose. This type of HMMs training can be suggested to everyone whose research projects or applications require Hidden Markov Models of speech.

1. Úvod

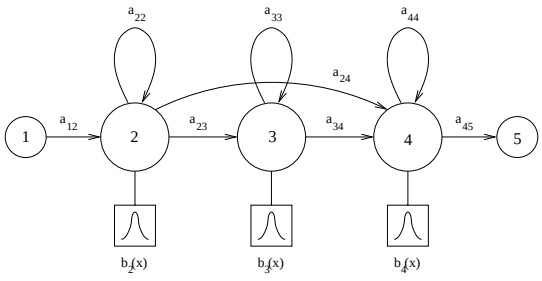
Systémy pro rozpoznávání a syntézu řeči (obecněji pro zpracování řečových signálů) jsou vyvíjeny a zkoumány po dobu několika desetiletí. Výzkum a vývoj v této oblasti je motivován i nemalým zájmem ze strany komerční sféry. Poměrně rychlý rozvoj výpočetní techniky umožňuje implementovat kvalitnější, avšak výpočetně náročnější algoritmy pro zpracování signálů. To je jeden z důvodů rozšíření metod zpracování řeči využívajících skrytých Markovových modelů (HMM). Algoritmy založené na skrytých Markovových modelech nacházejí hlavní uplatnění při rozpoznávání řeči, jejich další využití nalézáme při automatické tvorbě fonémových přepisů řečových nahrávek a při nejrůznějších experimentech v oblasti zpracování řečových signálů.

Pokud chceme využít HMM modelů pro nějakou aplikaci, případně experiment, jsme téměř vždy postaveni před problém získání vhodných HMM modelů. Tento článek si klade za cíl seznámit čtenáře se způsobem získávání HMM modelů hlásek (fonémů), případně kontextově závislých hlásek, ze středně rozsáhlé databáze obsahující několik desítek hodin řečových nahrávek získaných od různých mluvčích. Pro napsání tohoto textu bylo využito poznatků získaných při trénování kontextově závislých fonémů [5] pomocí databáze SpeechDat [4] a volně dostupného softwaru HTK (Hidden Markov Model Toolkit) [2].

2. Pojem Markovových modelů

Nejprve uvedeme Markovův model a zavedeme některé často užívané pojmy. Podrobnější a ucelenější výklad problematiky HMM aplikované na řečové signály lze nalézt v [1, 2, 3].

Skrytý Markovův model odpovídá statistickému modelu automatu s konečným počtem stavů, který může být užitečný pro popis některých nestacionárních signálů jako je například řeč. Jeho využití motivuje představa, že řečový signál je možné (za předpokladu určitých zjednodušení) považovat za sled po sobě následujících kvazistacionárních segmentů signálu se specifickými statistickými parametry. Tyto kvazistacionární segmenty signálu, případně jejich kombinace, mají význam jednotlivých úseků řeči jako je např. hláska, slovo apod. a reprezentují tak znaky nutné pro přenos informace. Pokud se podaří správně přiřadit jednotlivé kvazistacionární části řečového signálu jejich modelům (reprezentovaných jednotlivými stavy uvnitř HMM modelu), pak již dokážeme určit informační obsah obsažený v řečovém signálu.



Obrázek 1: Příklad pětistavového, levo-pravého skrytého Markovova modelu, kde první a poslední stav je neemitující.

Schéma příkladu skrytého Markovova modelu je znázorněno na obr.1. N-stavový HMM model je definován následujícím souborem parametrů $\lambda = \{\pi_i,\ a_{ij},\ b_i(x),\ i,\ j = 1, ..., N\}$, kde π_i označuje pravděpodobnost počátečního stavu, a_{ij} označuje pravděpodobnost přechodu mezi stavy, $b_i(x)$ označuje funkci hustoty pravděpodobnosti v daném stavu pro vektor pozorování x (ten je odvozen z řečového signálu), proměnné i a j označují konkrétní stavy v HMM modelu. Nejvýznamnějšími parametry HMM modelu jsou pravděpodobnosti přechodů mezi jednotlivými stavy a hustoty pravděpodobností v jednotlivých stavech. Pravděpodobnosti přechodů statisticky popisují doby trvání sledu kvazistacionárních segmentů signálu se stejnými statistickými parametry, hustoty pravděpodobnosti v jednotlivých stavech naopak vyjadřují statistické parametry těchto segmentů signálu. Vektor pozorování x, jehož statistické parametry vyhodnocujeme pomocí hustoty pravděpodobnosti, reprezentuje vždy daný segment signálu a je odvozen z řečového signálu pomocí procesu, který se nazývá parametrizace (viz. kapitola 3.2).

Funkce hustoty pravděpodobnosti, pomocí níž posuzujeme příslušnost pozorovaného kvazistacionárního segmentu signálu k danému stavu, má často tvar

$$b_i(x) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} c_{ism} N(x_s; \mu_{ism}, \Sigma_{ism}) \right]^{\gamma_s}, \qquad (1)$$

kde M_s je počet tzv. mixtures, S je počet tzv. streams (viz. kapitola 5.4), c_{ism} je váha každé m-té komponenty, $N(x_s; \mu_{ism}, \Sigma_{ism})$ je hustota pravděpodobnosti vícerozměrného normálního rozdělení s parametry μ_{ism} (vektor středních hodnot) a Σ_{ism} (kovarianční matice) a γ_s nastavuje váhu každého streamu s. Při volbě počtu mixtures a streams pro konkrétní aplikaci je nutné řešit kompromis mezi snahou o kvalitní vyjádření hustoty pravděpodobnosti v daném stavu a výpočetními nároky. Markovovy modely s takto definovanou hustotou pravděpodobnosti se také nazývají CDHMM (Continues Density Hidden Markov Model). Modelům, kde jsou přechody mezi jednotlivými stavy definovány pouze z leva do prava, se říká levopravé.

Takto definovaný HMM model může představovat frázi, slovo, foném, případně kontextově závislý foném. HMM pak popisuje z jakých segmentů řečového signálu se daný řečový úsek skládá (specifikuje doby trvání jejich sledu, statistické parametry a pořadí segmentů). Při rozpoznávání řeči je nutné umět vyjádřit s jakou pravděpodobností byl úsek řečového signálu s neznámým informačním obsahem vygenerován daným HMM modelem. Pokud je sekvence vektorů pozorování definována jako $X = [x_1, x_2, ..., x_n]$ (tato sekvence je odvozena ze segmentů řečového signálu), pak můžeme tuto pravděpodobnost vypočítat podle vzorce

$$P(X|M) = \sum_{U} a_{u(0)u(1)} \prod_{t=1}^{T} b_{u(t)}(x_t)\, a_{u(t)u(t+1)}, \qquad (2)$$

kde $U = u(1), u(2), u(3), ..., u(T)$ představuje všechny možné sekvence stavů, kterými mohl být signál generován. Z takto nalezené pravděpodobnosti lze přímo usoudit na to, který HMM model (představující například určité slovo) nejvíce odpovídá dané sekvenci pozorování. Takto lze získat informační obsah uložený v řečovém signálu, což je úlohou rozpoznávání řeči. Výpočet pravděpodobnosti podle předchozí definice je velice výpočetně náročný, proto se při reálných aplikacích používá efektivnějších způsobů výpočtu [2, 3].

Jak již bylo řečeno, před vlastním využitím HMM modelů, které popisují úseky řečového signálu, je nutné tyto modely (resp. jejich parametry) nejdříve získat. Nejčastěji se tak děje procesem nazývaným trénování. Před trénováním nejdříve definujeme topologii modelu (počet stavů, přechody mezi nimi apod.), a pak, pomocí databáze řečových signálů, ke kterým známe jejich informační obsah, upřesňujeme parametry modelu (pravděpodobnosti přechodů a parametry hustot pravděpodobností jednotlivých stavů). Často používaným algoritmem pro odhad parametrů modelu je tzv. Baum-Welchova reestimace [2, 3].

Další část textu je zaměřena na popis získávání HMM modelů fonémů a kontextově závislých fonémů ze středně rozsáhlé databáze řečových signálů. Modely fonémů jsou oblíbeny pro jejich malý počet a pro skutečnost, že HMM modely slov lze složit z modelů fonémů. Pro přímé získání kvalitních HMM modelů slov, je nutné mít poměrně rozsáhlou databázi řečových nahrávek těchto slov získaných od různých mluvčích. HMM model slova získaný z HMM modelů jednotlivých hlásek není vždy dostatečně kvalitní pro konkrétní aplikaci. To je způsobeno tím, že výslovnost fonému je ovlivněna fonémem, který ho předchází a následuje. Tato skutečnost motivuje vytváření tzv. kontextově závislých HMM modelů fonémů, které se snaží tento problém řešit[1].

3. Příprava databáze

K trénování HMM modelů musíme mít k dispozici vhodnou databázi nahrávek řečových signálů i s popisem jejich obsahu. Pokud chceme získat kvalitní HMM modely fonémů, které jsou nezávislé na konkrétním mluvčím, pak je nutné použít databázi v které jsou uloženy nahrávky získané minimálně od několika stovek mluvčích. Vytváření takové databáze je časově a finančně náročná činnost. Pokud chceme trénovat kontextově závislé fonémy, je třeba navíc dbát o tzv. fonémovou vyváženost databáze. Zde byla při trénování využita telefonní databáze SpeechDat [4].

Organizace každé konkrétní databáze je odlišná. Aby však daná databáze byla pro náš účel použitelná, musí obsahovat dva typy souborů. Prvním typem jsou soubory s vlastními nahrávkami řečových signálů. Druhý typ souborů je často uložen v textovém tvaru a obsahuje informace, co se v dané nahrávce vyskytuje za slova, případně hlásky nebo věty a některé další doplňkové informace jako například věk řečníka, typ nahrávacího zařízení apod. Oba zmíněné typy souborů je obvykle nutné postupně zkonvertovat do formátu vhodného pro trénování HMM modelů.

3.1. Vytvoření labelovacích souborů

První částí konverze je převedení souborů do formátu, který je schopen načíst vybraný trénovací software (např. HTK). Po tomto převedení celé databáze získáme soubory s nahrávkami ve vhodném formátu pro HTK a tzv. labelovací soubory, ve kterých jsou v textovém tvaru zapsána slova, která daný soubor s nahrávkou obsahuje.

Jestliže chceme trénovat modely fonémů a ne celých slov, pak je nutné labelovací soubory na úrovni slov nahradit labelovacími soubory na úrovni fonémů, tzn. nahradit slova jejich fonetickým přepisem. Nejdříve je nutné zavést

[1] Počet fonémů je v každém jazyce jiný a pohybuje se od 12 do 60, počet kontextově závislých fonémů teoreticky dosahuje třetí mocniny počtu fonémů.

sadu fonémů, která bude odpovídat názvům později vzniklých HMM modelů těchto fonémů. Zde byla využita sada fonémů dodaná k databázi SpeechDat [4], kde konkrétní názvy fonémů byly změněny tak, aby byly bez problémů zpracovatelné softwarem HTK. Výpis názvů fonémů pak může být např. následující

```
a, aa, au, b, tt, e, ee, sil, sp  ...apod.,
```

kde `aa` označuje á, `tt` označuje ť apod. Zvláštní význam mají názvy `sil` a `sp`, kde `sil` je název modelu pro „dlouhou“ pauzu v řeči a `sp` je název modelu pro „krátkou“ pauzu v řeči.

Po zavedení sady fonémů potřebujeme získat slovník s výslovností jednotlivých slov obsažených v databázi, kde je jejich výslovnost zapsána pomocí takto definované sady. Tento slovník může být součástí databáze, tak je tomu např. u databáze SpeechDat, nebo ho lze, v případě českého jazyka, vytvořit syntézou výslovnosti podle lingvistických pravidel. Pro ilustraci je zde uvedena malá část takového slovníku:

```
abecední a b e c e d nn ii sp
absence a p s e n c e sp
absolventské a p s o l v e n t s k ee sp
absolvovala a p s o l v o v a l a sp ...atd.
```

Prvním výrazem na řádce je vždy dané slovo, pak následuje řada fonémů reprezentující dané slovo. Jako poslední foném je zde umístěn model pro „krátkou“ pauzu. Pokud byla definována sada fonémů a existuje slovník s výslovností jednotlivých slov pomocí těchto fonémů, pak je možné převést stávající labelovací soubory na úrovni slov na labelovací soubory na úrovni fonémů. Tato akce, jako i mnohé další, jsou podporovány softwarem HTK. Z důvodů, které budou zmíněny v další kapitole, jsou generovány dva typy labelovacích souborů na úrovni fonémů. První z nich neobsahuje modely pro „krátkou“ pauzu na konci jednotlivých slov, druhý je naopak obsahuje. Oba dva typy obsahují modely pro „dlouhou“ pauzu umístěnou na začátku a konci každé nahrávky. Pro ilustraci je opět uveden krátký výpis takto vzniklého labelovacího souboru, ve tvaru používaném v HTK.

```
#!MLF!#
"*/a30000a1.lab"
sil
cc
e
ss
tt
i
n
a
sil
.
```

Soubor uvádí krátká hlavička, v uvozovkách je (až na koncovku) zapsán název souboru s nahrávkou, ke které labelovací soubor patří, následuje výpis fonémů a tečkou je označen konec tohoto souboru.

3.2. Parametrizace nahrávek

Posledním krokem při přípravě dat pro trénování je jejich parametrizace tzn. převedení signálu z časové oblasti na posloupnost parametrizačních vektorů (vektorů pozorování viz. kapitola 2) v čase. HTK podporuje více druhů parametrizačních vektorů. Mezi často používané patří tzv. melovské kepstrální koeficienty s odvozenými delta a delta-delta koeficienty. Postup kódování řečového signálu do melovských kepstrálních koeficientů lze zkráceně popsat takto: Digitalizovaný signál v časové oblasti je rozložen na po sobě jdoucí řadu segmentů, kdy každý segment obsahuje např. 256 vzorků signálu. Z každého segmentu je vypočítáno jeho amplitudové (popř. výkonové) spektrum pomocí rychlé Fourierovy transformace (FFT). Takto vzniklé spektrální čáry jsou váhovány a vzájemně sčítány do menšího počtu pomocí tzv. melovské banky filtrů, která reprezentuje nelineární frekvenční osu (obdoba vnímání řeči sluchem). Amplitudy melovských spektrálních čár jsou dále logaritmovány a pomocí inverzní diskrétní kosinové transformace (IDCT) je vypočítáno výsledné kepstrum.

Při takto volené parametrizaci je opět nutné řešit kompromis mezi počtem kepstrálních koeficientů použitých v dané aplikaci, kdy větší počet lépe popisuje řečový signál a výpočetním výkonem, který je k dispozici. Často používaným kompromisem je volba třinácti kepstrálních koeficientů a stejného počtu delta a delta-delta (tzv. akceleračních) koeficientů. Delta kepstrální koeficienty jsou úměrné rozdílu odpovídajících si kepstrálních koeficientů získaných ze dvou po sobě jdoucích segmentů. Akcelerační koeficienty jsou získány obdobným způsobem, ale z delta kepstrálních koeficientů. Tyto koeficienty zachycují časový vývoj parametrů řeči.

4. Trénování HMM modelů fonémů

V této kapitole je popsán postup při trénování HMM modelů fonémů. Vstupem do celého procesu jsou soubory s parametrizovanými řečovými nahrávkami a soubory s fonetickým přepisem těchto nahrávek (labelovací soubory na úrovni fonémů). Počátečním bodem při trénování je sada HMM modelů fonémů s identickými parametry. Tyto modely jsou přetrénovány pomocí Baum-Welchova reestimačního algoritmu. K takto natrénovaným modelům je přidán model pro „krátké“ pauzy v řeči a dále je modifikován model pro „dlouhé“ pauzy v řeči. Takto modifikovaný soubor modelů je dále přetrénován.

Některá slova ve slovníku mají více možných výslovností. Když byly vytvářeny labelovací soubory pro jednotlivé nahrávky, byla vybrána vždy první z nich. Jestliže již máme natrénován soubor HMM modelů fonémů s „rozumnými“ parametry, pak můžeme využít softwaru HTK k dalšímu upřesnění fonetického přepisu. Tzn. najít která z možných výslovností danému slovu nejvíce odpovídá.

HMM modely jsou pak znovu přetrénovány s upřesněnými labelovacími soubory nahrávek.

4.1. Vytvoření prototypu HMM modelu fonému

Prvním krokem při trénování HMM modelů je definice jejich topologie. Při trénování HMM modelů fonémů je vhodné použít pro jednotlivé fonémy stejné levo-pravé modely se třemi stavy, bez přeskoků jednotlivých stavů. Takto vytvořený model, ve kterém prozatím nezáleží na jeho vnitřních hodnotách se nazývá prototypem. Následuje příklad reálné definice prototypu používané softwarem HTK (jedná se o klasický textový soubor).

```
~o <VecSize> 39 <MFCC_0_D_A>
~h "proto"
<BeginHMM>
 <NumStates> 5
 <State> 2
    <Mean> 39
      0.0 0.0 0.0 0.0 0.0 0.0 0.0 ...
    <Variance> 39
      1.0 1.0 1.0 1.0 1.0 1.0 1.0 ...
 <State> 3
    <Mean> 39
      0.0 0.0 0.0 0.0 0.0 0.0 0.0 ...
    <Variance> 39
      1.0 1.0 1.0 1.0 1.0 1.0 1.0 ...
 <State> 4
    <Mean> 39
      0.0 0.0 0.0 0.0 0.0 0.0 0.0 ...
    <Variance> 39
      1.0 1.0 1.0 1.0 1.0 1.0 1.0 ...
<TransP> 5
0.0 1.0 0.0 0.0 0.0
0.0 0.6 0.4 0.0 0.0
0.0 0.0 0.6 0.4 0.0
0.0 0.0 0.0 0.7 0.3
0.0 0.0 0.0 0.0 0.0
<EndHMM>
```

Takto definovaný HMM model obsahuje tři stavy. V souboru je sice definováno stavů pět, ale první a poslední jsou neemitující a neobsahují žádné definice hustot pravděpodobností. Definice každého stavu je uvozena klíčovým slovem `<State>`. Za tímto slovem následuje výpis středních hodnot a kovarianční matice hustoty pravděpodobnostního rozdělení daného stavu. Kovarianční matice je zde nahrazena vektorem hodnot její hlavní diagonály (variance), z důvodu snížení výpočetních nároků. Parametrizační vektor použitý v modelu má 39 prvků. Toto číslo je dáno součtem 12 kepstrálních koeficientů s nultým kepstrálním koeficientem vyjadřujícím energii segmentu (definován typem parametrizace `MFCC_0`), dále je přičteno 13 delta a 13 delta-delta kepstrálních koeficientů. Poslední částí definice je tzv. matice přechodů. V této matici jsou definovány pravděpodobnosti přechodů mezi jednotlivými stavy. Například pravděpodobnosti přechodu mezi stavem 2 a stavem 3 (a_{23}) odpovídá číslo umístěné na druhém řádku a třetím sloupci.

U takto definovaného modelu (napsaného např. pomocí textového editoru) je nutné alespoň orientačně určit jeho střední hodnoty a variance. K tomuto účelu software HTK disponuje funkcí, která projde parametrizované řečové soubory v databázi a pomocí těchto dat vypočítá globální střední hodnoty a variance v jednotlivých stavech. Takto získané hodnoty uloží do nově vzniklého prototypu.

4.2. Přetrénování HMM modelů fonémů

Před trénováním jednotlivých modelů fonémů je nutné nejdříve vytvořit soubor s definicemi těchto fonémů. Tento soubor vznikne jednoduše naklonováním průměrného modelu fonému a přepsáním jména pro každý foném. Takto dostaneme soubor nenatrénovaných HMM modelů fonémů.

Software HTK obsahuje dva programy vhodné pro trénování HMM modelů fonémů. První z nich je určen pro trénování HMM modelů z databází, kde kromě informačního obsahu nahrávek jsou k dispozici i časové značky, které označují výskyt signálů, na které se mají HMM modely natrénovat. Tento program využívá ke své činnosti tolik iterací Baum-Welchovy reestimace, dokud daný model nezkonverguje.

Protože takovou databázi pro trénování modelů fonémů k dispozici nemáme, je využíván program pro tzv. vložené (embedded) trénování. Program pracuje následovně. Na počátku načte kompletní soubor definic HMM modelů. Ke každému souboru s nahrávkou musí být připojen odpovídající labelovací soubor. Pomocí labelovacího souboru na úrovni fonémů a modelů fonémů se vytvoří jeden kompozitní (složený) HMM model představující celou jednu nahrávku (zpravidla čítající několik slov). Takto vzniklý model je přetrénován jednou iterací Baum-Welchova algoritmu a tím jsou zároveň reestimovány i modely, ze kterých je kompozitní HMM model složen. Takto je postupně zpracována celá databáze řečových nahrávek a výsledné modely fonémů jsou vypočítány pomocí vážených průměrů. Vzhledem k tomu, že program pracuje pouze s jednou iterací Baum-Welchova algoritmu, je vhodné reestimaci modelů dvakrát nebo i vícekrát opakovat.

Při tomto prvním trénování ještě není možné přesně odlišit modely fonémů a modely pro krátké pauzy. Z tohoto důvodu jsou použity labelovací soubory, kde modely pro krátké pauzy zahrnuty nejsou. Modely pro dlouhé pauzy jsou umístěny na začátku a konci každé nahrávky, kde je jistota, že se pauzy opravdu vyskytují.

4.3. Úprava modelů pro pauzy v řeči

V předcházejících krocích byly vytvořeny třístavové levo-pravé HMM modely pro každý foném a jeden model pro „dlouhou“ pauzu s názvem `sil` se stejnou strukturou. Do modelu pro „dlouhou“ pauzu jsou v dalším kroku přidány přechody ze stavu 2 do stavu 4 a zpátky. Tato modifikace

je provedena proto, aby se stal tento model univerzálnější pro různé typy šumů. Zpětný přechod ze stavu 2 do stavu 4 umožňuje zůstat uvnitř modelu pauzy bez přechodu do dalšího slova.

Zcela nově je vytvořen model pro „krátkou“ pauzu s názvem `sp`. Tento model má pouze jeden stav, který je svázán se stavem 3 modelu `sil` (tzn. že tento stav má stejné parametry) a také obsahuje přechod ze svého vstupu přímo na výstup.

Po těchto úpravách je opět alespoň dvakrát spuštěna procedura pro reestimaci parametrů HMM modelů. Jediným rozdílem oproti trénování popisovaném v podkapitole 4.2 je použití labelovacích souborů obsahujících kromě názvů jednotlivých fonémů také název pro model „krátkého“ ticha, ten je pak umístěn vždy na hranici jednotlivých slov.

4.4. Zarovnání trénovacích dat

Slovník s fonetickým přepisem jednotlivých slov obsahuje pro některá slova více možných výslovností. Když byly vytvářeny labelovací soubory pro jednotlivé nahrávky byla vybrána vždy první z nich. Modely fonémů, které byly v předchozím postupu vytvořeny, mohou být využity k vytvoření nových (přesnějších) labelovacích souborů na úrovni fonémů.

Pro dané slovo s více možnými výslovnostmi se vytvoří všechny možné kompozitní HMM modely, které slovník nabízí. Jako správná výslovnost, patřící k realizaci daného slova v nahrávce, je pak vybrána ta, jejíž kompozitní model odpovídá dané nahrávce s nejvyšší pravděpodobností. Takto nově vygenerované výslovnosti, zapsané pomocí fonémů, jsou uloženy do nových labelovacích souborů.

Nyní lze modely fonémů s pomocí upřesněných labelovacích souborů znovu přetrénovat a získat tak výsledné modely fonémů. Tyto fonémy jsou již prakticky použitelné např. pro konstrukci jednoduchého rozpoznávače povelů.

5. Trénování HMM modelů kontextově závislých fonémů

V předchozí kapitole byly získány HMM modely pro jednotlivé fonémy. S těmito modely je již možné sestavit např. rozpoznávač povelů, případně i jednoduchých vět. Lepších výsledků lze však dosáhnout s tzv. kontextově závislými HMM modely fonémů. Tvorba modelů kontextově závislých fonémů je motivována skutečností, že výslovnost daného fonému a tím i parametry jeho modelu se liší podle toho, který foném ho předchází a který následuje. Pro každou takovouto kombinaci je vytvořen zvláštní model (tzv. kontextově závislý), kterému se říká trifón.

Trénování HMM modelů trifónů se děje ve dvou krocích. Nejprve se fonémový přepis nahrávek převede na trifónový a soubor modelů trifónů se vytvoří naklonováním modelů fonémů a následným přetrénováním. Následně jsou jednotlivé stavy s podobnými akustickými parametry sjednoceny (svázány) s cílem stabilnějšího odhadu jejich parametrů.

5.1. Změna fonémového přepisu nahrávek na trifónový

Prvním krokem, který je nutné provést před vlastním trénováním modelů trifónů je převod fonémových labelovacích souborů na trifónové labelovací soubory. Tento převod lze uskutečnit velice snadno, protože je podporován softwarem HTK. Následuje příklad takového labelovacího souboru pro nahrávku slova „čeština“.

```
#!MLF!#
"*/a30000a1.lab"
sil
cc+e
cc-e+ss
e-ss+tt
ss-tt+i
tt-i+n
i-n+a
n-a
sp
sil
.
```

Jak lze snadno vydedukovat z předcházejícího příkladu, název modelu trifónu se skládá z názvu fonému, který předchází (oddělen znaménkem mínus), z původního názvu fonému a z názvu fonému, který následuje (oddělen znaménkem plus). Na začátku a konci slova vznikají tzv. difóny. Při automatickém generování nových labelovacích souborů je vhodné také vytvořit seznam všech kontextově závislých fonémů (difónů a trifónů), které se v databázi alespoň jednou vyskytují.

5.2. Změna HMM modelů fonémů na kontextově závislé modely fonémů

Před trénováním modelů trifónů je nejdříve nutné vytvořit jejich modely. Tyto modely vzniknou naklonováním původních modelů fonémů. Generování modelů trifónů se provádí automaticky podle seznamu trifónů vyskytujících se alespoň jednou v databázi. Pro každý model trifónu ve tvaru `a-b+c` uvedený v seznamu se najde model fonému `b` a vytvoří se jeho kopie s názvem `a-b+c`. Ne každý trifón se v databázi vyskytuje v dostatečném množství pro natrénování jeho kvalitního modelu (tato problematika bude podrobněji diskutována v příští podkapitole). Z tohoto důvodu jsou sjednoceny matice přechodů u modelů trifónů, odvozených od stejného modelu fonému. V praxi to znamená, že např. u všech modelů trifónů odvozených od modelu `a` bude matice přechodů pouze jedna a modely ji budou sdílet.

Jestliže máme vytvořeny trifónové přepisy nahrávek a definice HMM modelů trifónů i s jejich seznamem, pak je

vše připraveno k trénování HMM modelů trifónů. Toto trénování se děje obdobně jako při trénování modelů fonémů.

5.3. Vytvoření HMM modelů trifónů s vázanými parametry jednotlivých stavů

Předešlým postupem lze natrénovat sadu trifónů, kde vždy trifóny odvozené od daného fonému sdílejí společnou matici přechodů. Při odhadu parametrů HMM modelů trifónů často nastává stav, kdy pro velké množství variancí z pravděpodobnostních rozdělení jednotlivých stavů není možné získat kvalitní odhad z důvodu nedostatku trénovacích dat pro daný stav. Tento nedostatek lze řešit pomocí vázání a následného sdílení parametrů jednotlivých stavů mezi jednotlivými trifóny. Tímto způsobem je zajištěn stabilnější odhad statistických parametrů (středních hodnot a variancí) jednotlivých stavů.

Vázání parametrů jednotlivých stavů však vyžaduje větší přesnost než tomu bylo u matic přechodů, protože na správnosti odhadu těchto parametrů značně záleží kvalita výsledných modelů. Software HTK podporuje dva způsoby shlukování (clustering) a následného vázání parametrů jednotlivých stavů pomocí maker. První z nich je řízen daty a využívá míru podobnosti mezi jednotlivými stavy. Druhá metoda využívá ke své činnosti tzv. rozhodovacího stromu. Ten je založen na zjišťování levého a pravého kontextu každého trifónu. Tyto rozhodovací stromy se snaží nalézt ty kontexty, které způsobují největší rozdíly mezi akustickou výslovností a proto by měly rozdělovat jednotlivé shluky stavů.

Rozhodovací stromy jsou vytvářeny pomocí sady otázek, pomocí kterých se stavy u specifikované třídy fonémů rozdělují do shluků. Každý shluk je nakonec svázán do podoby makra, které zastupuje jeho parametry. Jednotlivé otázky jsou definovány sadou kontextů. Jako příkladem realizace otázky v HTK může posloužit výpis `QS "R_nasal1" {*+m,*+n,*+nn,*+mv,*+ng}`. Tato otázka má jméno `R_nasal1` a je pravdivá, jestliže pravý kontext daného trifónu odpovídá fonémům `m, n, nn, mv` nebo `ng`. U HMM modelů trifónů je nutné definovat otázky odkazující jak na levý, tak i na pravý kontext trifónu. Při definici otázek je nutné využít i určitých lingvistických znalostí. Není však na škodu vytvořit otázky u kterých si nejsme úplně jisti, protože ty, které budou vyhodnoceny jako neužitečné budou ignorovány.

Po vygenerování modelů trifónů s vázanými parametry jednotlivých stavů je vhodné tuto sadu opět několikrát přetrénovat za účelem stabilnějšího a přesnějšího odhadu parametrů modelů. Výsledkem jsou pak již poměrně kvalitní modely kontextově závislých fonémů.

5.4. Streamy a mixtures

Až doposud bylo k popisu hustot pravděpodobnostních rozdělení jednotlivých stavů používána jedna hustota odpovídající vícerozměrnému normálního rozdělení. Kvalitu HMM modelů pro popis řeči lze zvýšit úpravou výpočtu hustot pravděpodobností v jednotlivých stavech. Tato úprava spočívá v rozdělení parametrizačního vektoru na např. tři nezávislé vektory (streams, odpovídající statickým kepstrálním, delta kepstrálním a akceleračním koeficientům) a pro každý tento vektor počítat zvlášť hustotu pravděpodobnosti jako součet např. tří hustot pravděpodobnosti s vícerozměrným normálním rozložením (mixtures). Vynásobením hustot pravděpodobnosti pro jednotlivé nezávislé vektory (streams) získáme hustotu pravděpodobnosti daného stavu.

Software HTK umožňuje přidat mixtures a streams do definic HMM modelů dodatečně. Počet mixtures je volitelný a lze ho navrhnout podle toho, jaký výpočetní výkon bude v konkrétní aplikaci k dispozici. Počet streams je rovněž volitelný, i když ne zcela libovolně. Celkem logickým se jeví rozdělení parametrizačního vektoru na části, které k sobě náleží způsobem výpočtu (např. statické, delta, delta-delta kepstrální koeficienty). Po úpravě HMM modelů je nutné tyto modely opět několikrát přetrénovat.

6. Závěr

V článku byl popsán moderní způsob trénování skrytých Markovových modelů řeči užívaný v poslední době. Popis byl zaměřen na trénování HMM modelů fonémů a kontextově závislých fonémů, vzhledem k jejich aplikačním výhodám. Tento způsob trénování HMM modelů nevyžaduje použití databáze s ručně vytvořenými časovými značkami pro popis databáze, stačí pouze popis na úrovni slov, a proto ho lze vřele doporučit všem zájemcům, kteří pro své výzkumy nebo aplikace vyžadují kvalitní HMM modely řeči.

Poděkování

Tato práce byla podpořena grantem GAČR č.102/02/0124 „Hlasové technologie v podpoře informační společnosti“.

Reference

[1] Rabiner, L. and Juang, B.-H.: *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[2] Young, S.: *The HTK Book (for HTK Version 3.1)*, Cambridge University Engineering Department, December 2001.

[3] Psutka, J.: *Komunikace s počítačem mluvenou řečí*, Academia, 1995.

[4] Černocký, J., Pollák, P., Hanžl, V.: *SpeechDat(E) Czech Database for the Fixed Telephone Network*, VUT and ČVUT, October 2000.

[5] Novotný, J.: *Trénování a využití kontextově závislých HMM modelů fonémů*, Výzkumná zpráva, katedra teorie obvodů ČVUT, Srpen 2002.

Akustické listy: ročník 8, číslo 3 září 2002 ISSN: 1212-4702
Vydavatel: Česká akustická společnost, Technická 2, 166 27 Praha 6 Vytisklo: Ediční středisko ČVUT
Počet stran: 24 Počet výtisků: 200
Číslo připravili: Marek Brothánek, Ondřej Jiříček, Jan Kozák

Příspěvky nejsou redakčně upravovány. Za jazykovou úpravu odpovídají jejich autoři.
Uzávěrka příštího čísla Akustických listů je 29. listopadu 2002.