VÝVOJ SYSTÉMU AUTOMATICKÉHO ROZPOZNÁVANIA REČI
DEVELOPMENT OF A SYSTEM FOR AUTOMATIC RECOGNITION OF SPEECH

Roman Jarina, Michal Kuba
Katedra telekomunikácií, Elektrotechnická fakulta, Žilinská univerzitaVeľký diel, 010 26 Žilina
E-mail: {jarina, kuba}@felutc.sk

Abstrakt Článok podáva prehľad výskumu v oblasti spracovania a automatického rozpoznávania rečových signálov (ARR) na Katedre telekomunikácií Elektrotechnickej fakulty Žilinskej univerzity. Súčasný výskum je zameraný na parametrizáciu reči pomocou dvojrozmernej kepstrálnej analýzy a na aplikáciu Skrytých Markovovych modelov a neuronových sietí pre rozpoznávanie slov v slovenskom jazyku. Článok stručne zhrňuje dosiahnuté výsledky a naznačuje budúcu orientáciu výskumu v oblasti ARR.

Summary The article gives a review of a research on processing and automatic recognition of speech signals (ARR) at the Department of Telecommunications of the Faculty of Electrical Engineering, University of Žilina. On-going research is oriented to speech parametrization using 2-dimensional cepstral analysis, and to an application of HMMs and neural networks for speech recognition in Slovak language. The article summarizes achieved results and outlines future orientation of our research in automatic speech recognition.

1. ÚVOD

Výskum v oblasti číslicového spracovania rečových signálov sa začal na katedre telekomunikácií Žilinskej univerzity len nedávno. Počiatočné snahy boli orientované na analýzu a kódovanie rečových signálov [1], [2]. Toto úsilie vyústilo do úspešnej realizácie rýchleho parametrického kódera reči. Tento kóder pracuje v reálnom čase a dosahuje priemernú dátovú rýchlosť kódovaného signálu 3,2 – 4,2 kbit/s [2].

S nástupom nových komunikačných a infor-mačných technológií vznikla naliehavá potreba riešenia komunikácie človeka s počítačom (informačným systémom) prirodzenou ľudskou rečou. Aj keď v oblasti rozpoznávania a syntézy reči sa vo svete dosiahli významné pokroky a týmto problémom sa venujú mnohé významné inštitúcie, vyvinuté dialógové rečové systémy stále ani zďaleka nedosahujú kvalitu prirodzenej ľudskej komunikácie. A navyše prevážna väčšina výskumu sa orientuje na rozpoznávanie a syntézu reči v anglickom jazyku. Menšinové jazyky, ku ktorým patrí aj slovenčina, sú vo veľkej nevýhode, pretože malé krajiny nedisponujú dostatočným ľudským a finančným potenciálom na riešenie tejto zložitej úlohy a nemôžu samozrejme konkurovať anglicky hovoriacim krajinám. Preto na katedre telekomunikácií vznikla myšlienka venovať sa tiež výskumu v oblasti analýzy a automatického rozpoznávania reči (ARR) v slovenskom jazyku.

Článok stručne zhrňuje naše dosiahnuté výsledky a naznačuje budúcu orientáciu výskumu v oblasti ARR.

Návrh systému ARR môžeme rozdeliť do dvoch častí:

a) Parametrizácia rečového signálu

b) Klasifikácia príznakov

Obidve časti návrhu sú rovnako náročné, a pretože výber a počet parametrov (príznakov) závisí aj od typu klasifikátora, nie je možné navrhovať tieto dve časti oddelene.

Z hľadiska klasifikácie príznakov (parametrov) reči je možné metódy rozpoznávania členiť do troch významných skupín: 1) metódy založené na porovnávaní vzorov, 2) metódy založené na štatistickom modelovaní produkcie reči, 3) metódy akusticko-fonetického dekódovania s uplatnením znalostných princípov produkcie reči.

Systém ARR vyvíjaný na katedre telekomunikácií využíva štatistické princípy založené na učiacich algoritmoch. Prvý variant klasifikuje príznaky reči pomocou neurónovej siete (NS) a druhý variant, ktorého vývoj je ešte len v počiatočnej teoretickej fáze, využíva skryté Markovove modely (HMM) [3].

2. PROBLÉMY SPOJENÉ S ARR

Ľudská reč je tvorená časovou postupnosťou zvukov. Pri dorozumievaní tak človek vníma zvuky, ktoré sa v určitých charakteristikách podobajú, a preto opakovane identifikuje v plynulej reči tie isté prvky. Preto spojitý rečový signál je možné rozložiť na elementárne časti, ktoré sú z hľadiska rozpoznávania reči určujúce. Týmito elementárnymi časťami, teda najmenšími akustickými jednotkami reči sú z jazykového hľadiska *hlásky* a z fonetického hľadiska *fonémy*. Fonéma tak reprezentuje celú škálu navzájom rôznych elementárnych zvukov, ktoré nesú ten istý informačný obsah a naviac v sebe zahŕňa celý proces artikulácie, ktorý je potrebný pre jej vznik.

Úloha rozpoznávania reči spočíva v schopnosti systému ARR identifikovať v rečovom signále

jednotlivé fonémy a automaticky rekonštruovať celý prednes na takej úrovni, aby na informácie vystupujúce zo systému ARR optimálne reagovali ďalšie nadväzujúce aplikačné systémy. Až tieto kombinácie systémov sú užitočné, lebo samostatný systém ARR bez náväznosti na ďalšie aplikácie nemá význam a nie je ani možné systém logicky a efektívne navrhnúť.

Aj napriek tomu, že sa v posledných rokoch učinili veľké pokroky v rozpoznávaní reči, konštrukcia systému, ktorý by bol schopný rozpoznať plynulý prednes od akéhokoľvek rečníka, ktorý používa ľubovoľné slová daného jazyka je vzdialenou budúcnosťou. Dôvody, ktoré rozpoznávanie reči stále veľmi komplikujú, je možné zhrnúť v týchto bodoch:

- Veľké odlišnosti hlasu medzi hovoriacimi (parametre hlasového ústrojenstva, spôsob artikulácie) ako aj v rámci jedného hovoriaceho (vplyv stresu, nálady).
- Meniace sa akustické pozadie, t.j. prítomnosť šumu a hluku môže spôsobiť veľké problémy pri rozpoznávaní reči.
- Vplyv prenosovej charakteristiky elektroakustických meničov a prenosových ciest.
- Kontextová závislosť foném, t.j. tie isté fonémy sú vyslovované rozdielnym spôsobom v závislosti od predchádzajúcich vyslovených foném (hlások).
- Veľkosť slovnej zásoby: Počet prípustných slov je príliš veľký pre väčšinu aplikácií.
- Variabilita dialógov a prejavu: Tá istá myšlienka môže byť vyjadrená rozdielnymi slovami a to isté slovo môže byť vyjadrené pomocou rozdielnych foném.
- U prirodzenej súvislej reči absencia páuz medzi slovami a prítomnosť mimovoľných zvukov.

Súčasné technológie ARR pre dosiahnutie optimálnej činnosti v dohľadnom čase väčšine uvedených problémov nedokážu čeliť, pretože uvedené problémy sú striktne spojené s aplikáciami a produktami ARR. To znamená, že ARR systém bude pracovať optimálne za predpokladu, že sa pripustia určité obmedzenia ako napríklad:

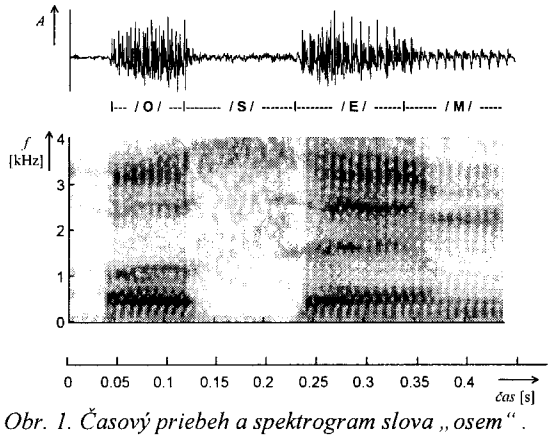
- rozpoznávanie izolovaných slov,
- rozpoznávanie reči len od jedného konkrétneho rečníka,
- obmedzený počet rozpoznávaných slov,
- úlohovo orientovaný systém ARR (t.j. jeho použitie sa obmedzí na niekoľko typov konštrukcií, ktoré budú riešiť menší počet síce jednoduchších konkrétnych úloh, ale zase o to spoľahlivejšie).

V súčasnosti obidva varianty nami navrhovaného systému ARR sú limitované na rozpoznávanie izolovaných slov. Veľkosť slovníka je obmedzená veľkosťou trénovacej databázy ktorú máme v súčasnosti k dispozícii (niekoľko desiatok slov v slovenskom jazyku od 15 rečníkov). V budúcnosti uvažujeme o rozšírení databázy v spolupráci so Slovenskou akadémiou vied.

3. PARAMETRIZÁCIA REČOVÉHO SIGNÁLU A 2-D KEPSTRÁLNA

V súčasnosti je už veľmi dobre známe, že spektrálne prechody (t.j. prechody medzi fonémami) hrajú veľmi dôležitú úlohu vo vnímaní reči, avšak úloha spektrálnych príznakov pri identifikácii reči z hľadiska popísania tejto dynamickej informácie nie je stále dostatočne objasnená. Mnohé experimenty potvrdili dôležitosť spektrálnych dynamických príznakov vo vnímaní reči a vedú k ich explicitnému modelovaniu v systémoch ARR [4]. Pre ilustráciu je na obr.1 znázornená časovo-frekvenčná závislosť signálu slova „osem". Zmeny v spektre pri prechode medzi hláskami (fonémami) sú dobre viditeľné.

V našom modeli je táto dynamická informácia (premenlivosť spektra v čase) vyjadrená pomocou tzv. *modulačného spektra* [5] s využitím dvojrozmernej kepstrálnej analýzy [6].



Obr. 1. Časový priebeh a spektrogram slova „osem".
Fig. 1. Speech waveform and spectrogram of the Slovak word „osem".

Dvojrozmerné (2-D) kepstrum vyjadruje príznaky reči v maticovom tvare, kde jeden rozmer predstavuje *kepstrum* (kosínusová transformácia z logaritmu spektra) a druhý rozmer vyjadruje zmenu jednotlivých kepstrálnych koeficientov v čase - tzv. *modulačné spektrum*. Teda 2-D kepstrum súčasne popisuje aj vzájomné vzťahy medzi susednými spektrálnymi vektormi.

Rečový signál je reprezentovaný postupnosťou kepstrálnych matíc, pričom analýza je vykonávaná po blokoch s konštantnou dĺžkou. Disriminačné vlastnosti takýchto príznakov reči sme skúmali na

automatickom rozlišovaní zvukovo podobných spoluhlások. Výsledky sú uvedené v [7].

4. ROZPOZNÁVANIE IZOLOVANÝCH SLOV POMOCOU NEUTRÓNOVEJ SIETE

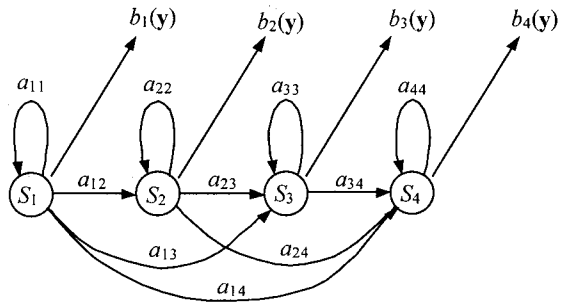
Rozpoznávanie predstavuje klasifikáciu vzorov (vektor príznakov) do príslušných tried. Počet tried je totožný s počtom rečových jednotiek (napr. slov) v slovníku. Každý vzor je tvorený súborom rečových príznakov, v našom prípade vektorom vytvoreným z vybraných koeficientov kepstrálnych matíc. Kritériom pre klasifikáciu môže byť vzdialenosť neznámeho vzoru od vzorov v knižnici. Rozpoznaná rečová jednotka bude potom tá, ktorej vzor v knižnici má najmenšiu vzdialenosť od neznámeho vzoru. Ako je ukázané v [8], výpočet vzdialenosti je veľmi náročný, pretože optimálny model musí zohľadňovať dôležitosť jednotlivých príznakov aj zo štatistického aj percepčného hľadiska. Preto sme pre klasifikáciu využili neurónovú sieť (NS). V tomto prípade nemusíme poznať presný matematický model, lebo NS je učená z príkladov.

Autor tohoto článku navrhol model, v ktorom je celé slovo vyjadrené v parametrickom tvare pomocou asi 200 príznakov vytvorených z troch kepstrálnych matíc. Touto metódou bola teda dosiahnutá výrazná redukcia dát oproti bežne používaným metódam (LPC, mel-kepstrum). Zníženie počtu príznakov (asi o jeden rád) umožnilo výrazne zjednodušiť štruktúru a tiež celý proces trénovania klasifikátora. Klasifikácia rečových príznakov je uskutočňovaná pomocou 3-vrstvovej a 4-vrstvovej NS s dopredným šírením (viacvrstvový perceptron s jednou resp. dvomi skrytými vrstvami). Na učenie siete bol použitý algoritmus spätného šírenia chyby s momentom a s adaptívnou rýchlosťou učenia. Model bol úspešne aplikovaný na rozpoznávanie slovenských čísloviek. Podrobný rozbor metódy s dosiahnutými výsledkami je v [8], [9].

5. ŠTATISTICKÝ PRÍSTUP K ROZPOZNÁVANIU REČI S VYUŽITÍM TEÓRIE HMM

Štatistické metódy sú založené na modelovaní rečovej produkcie najmä s využitím teórie skrytých Markovových modelov (HMM = Hidden Markov Model). Celý rečový prednes môže byť teoreticky modelovaný jedným skrytým Markovovým modelom. V praxi sa však ako celok modelujú jednotlivé slová jedným HMM, alebo sú vytvárané Markovove modely subslovných jednotiek (napr. skupín foném ako napr. slabík, jednotlivých foném, alofón, prípadne jednotiek menších ako fonéma a pod.) a slovo je modelované zreťazením týchto subslovných modelov. Pre každú triedu rovnakých slov zo slovníka sú v procese trénovania určené parametre modelu a neznáme slovo je klasifikátorom zaradené do tej triedy, ktorú model produkuje s najväčšou pravdepodobnosťou.

Tzv. modely zľava doprava sa ukázali ako veľmi výhodné pre modelovanie udalostí odohrávajúcich sa v čase. Ich základnou vlastnosťou je, že proces začína príchodom prvého akustického pozorovania (tvoreného vektorom príznakov rečového signálu) z počiatočného akustického stavu modelu a so vzrastajúcim časom dochádza k prechodom zo stavov s nižšími indexmi do stavov s vyššími indexmi, alebo dochádza k zotrvaniu v tom istom stave.



Obr. 2. Príklad 4-stavového ľavo-pravého HMM.
Fig. 2. An example of the 4-state left-right HMM.

Príklad 4-stavového skrytého Markovovho modelu zľava doprava je uvedený na obr. 2. Krúžky predstavujú jednotlivé akustické stavy a šípky vyjadrujú možné prechody medzi stavmi s označením príslušných pravdepodobností prechodov, ktoré sa tiež nazývajú tranzitné pravdepodobnosti. V každom časovom okamihu sa model nachádza v jednom zo štyroch stavov. V každom stave je daná hustota rozdelenia pravdepodobnosti výskytu daného pozorovania. Táto hustota pravdepodobnosti (emisná pravdepodobnosť) v podstate oceňuje možnosť výskytu daného akustického pozorovania v danom stave v danom čase. Tieto rozdelenia sú vo všeobecnosti rôzne pre každý jednotlivý stav a pre každý stav toto rozdelenie nezávisí od času.

Ak daná neznáma postupnosť akustických pozorovaní, ktorej chceme priradiť lexikálny význam (napr. slovo) s najvyššou pravde-podobnosťou vyhovuje niektorému HMM, ktoré sú vytvorené napr. pre každé slovo zo slovníka, potom tento HMM určuje to slovo, ktoré je možné považovať za rozpoznané. Tým je úloha rozpoznávania reči vykonaná, pretože bolo vykonané priradenie lexikálneho významu neznámemu slovnému prednesu reprezentovaného postupnosťou príznakov resp. akustických pozorovaní. Spomenutá pravdepodobnosť, ktorá oceňuje to, ako vyhovuje neznáma postupnosť akustických pozorovaní niektorému HMM sa nazýva vierohodnosť pozorovania.

Výhoda HMM spočíva v tom, že dobre modelujú premenlivosti v dĺžke trvania prednesu. Priamy

prechod modelom predstavuje priemernú dĺžku trvania prednesu, zotrvanie v niektorom stave odpovedá predĺženiu a preskočenie niektorého stavu či stavov naopak predstavuje skrátenie dĺžky trvania prednesu.

Podobne ako pri iných metódach rozpoznávania aj v tomto prípade sa určujú parametre HMM v procese trénovania. Odhad parametrov modelu, t.j. odhad tranzitných a emisných pravdepodobností sa vykonáva na základe dostatočného množstva trénovacích príkladov toho istého rečového prednesu, pre ktorý má byť model vytvorený.

6. ZÁVER

Prvý variant systému pre rozpoznávanie izolovaných slov uvedený v kapitole 4 má niektoré nevýhody (napr. stráca sa informácia o dĺžke slov), ktoré obmedzujú jeho použitie v systémoch s väčším slovníkom [8]. Preto v druhom variante bude rečový signál vyjadrený väčším počtom matíc. Klasifikátor bude vytvorený pomocou HMM. Vo svete už boli skúmané možnosti klasifikácie 2-D kepstrálnych príznakov pomocou HMM na úrovni celých slov [10], [11]. V tomto prípade bol počet pozorovaní, t.j. počet kepstrálnych matíc, rovnaký ako počet rámcov, ktorých celková doba trvania odpovedala jednému slovu. Súčasný výskum na katedre telekomunikácií sa venuje zjednodušeniu uvedeného modelu tak, aby sa redukoval výsledný počet pozorovaní, pretože sa dá predpokladať, že 2-D kepstrálna analýza dokáže dobre postihnúť koartikulačné efekty pri vyslovovaní slov. Výhoda uvedeného prístupu spočíva v tom, že sa výrazným spôsobom redukuje počet potrebných výpočtových operácií na rozpoznanie. Na druhej strane sa predpokladá, že skrytý Markovov model v spojitosti s 2-D kepstrálnou analýzou dokáže štatisticky uchovať poznatky o variáciách vo veľkom množstve rôznych akustických realizácií toho istého slova.

LITERATÚRA

[1] Jarina, R.: Improvement of Pitch Determination Based on Autocorrelation Method, Zborník medzin. vedeckej konf. ELEKTRO 97, Žilina, jún 1997, s. 214-217.

[2] Kuba, M., Jarina, R.: Real-Time Implementation of speech coder for PC, Zborník medzin. vedeckej konf. ELEKTRO 99, Žilina, máj 1999, s.112-114.

[3] Kuba, M.: Systémy automatického rozpoznávania reči, Práce a štúdie Žilinskej univerzity, zv. 26, EDIS Žilina, 2000, s. 27-43.

[4] Hanson, B.A., Applebaum, T.H., Junqua, J.C.: Spectral Dynamics for Speech Recognition under Adverse Conditions. In Advanced Topics In Automatic Speech and Speaker Recognition, Lee C.H.- Paliwal K.K. and Soong F.K. (Eds), Kluwer Academic Publishers 1995, pp. 331-356.

[5] Hermansky, H.: The Modulation Spectrum in Automatic Recognition of Speech. In 1997 IEEE Workshop on Automatic Speech Recognition and Understanding, IEEE Signal Processing Society, editors S. Furui and B.-H. Juang and W, Chou, 1997.

[6] Ariki, Y., Mizuta, S., Nagata, M. and Sakai, T.: Spoken-Word Recognition Using Dynamic Features Analysed by Two-Dimensional Cepstrum, Proc. IEE, Vol. 136, Pt.I, No.2, April 1989, pp.133-140.

[7] Jarina, R.: "Study of Discriminative Properties of Two-Dimensional Cepstrum Analysis for Speech Recognition, Zborník medzin. česko-slovenskej vedeckej konf. RADIOELEKTRONIKA 99, Brno, apríl 1999, s. 168-171.

[8] Jarina, R.: Kepstrálno-spektrálny model pre rozpoznávanie rečových signálov, Doktorandská dizertačná práca, Žilinská univerzita, okt. 1999, Žilina.

[9] Jarina, R.: Neural Network as a classifier of Two Dimensional Cepstrum Analysis based Speech Features, Proc. of 4th Int. Conf. on Digital Signal Processing DSP'99, Herľany, Sept.1999, pp.118-120.

[10] Vaseghi, S.V., Conner, P.N., and Milner, B.P.: Speech Modelling using Cepstral-Time Feature Matrices and Hidden Markov Models, IEE Proc., Vol.140, Pt.I, No.5, Oct.1993, pp. 317-320.

[11] Jančovič, P., Macho, D., Nadeu, C., and Rozinaj, G.: Feature Selection in Cepstral-Time Matrices for Clean and Noisy Speech Recognition, Proc. Of 4th. Int. Workshop TEMPUS – TELECOMNET ITTW'98, Barcelona, July 1998, pp. 28-36.