Rozpoznávání samohlásek

Filip Orság, student 3.r. 2.st. magisterského studia,
Zpracováno na UIVT FEI VUT Brno
Školitel: Doc. Ing. František Zbořil, CSc.

Vowel recognition

Abstract: this article describes how to recognize the vowels in a continuous speech. There are discussed various types of neural networks that could be used to the vowel recognition.

Oblast zpracování řeči a neuronových sítí není v dnešní době novinkou, ale každopádně jsou obě tyto oblasti teprve v počátku svého vývoje, jehož cílem je vytvořit „myslící“ a především rozumným způsobem komunikující stroj. Hlavním problémem vývojových týmů dnešní doby je vytvoření uživatelsky co nejpřívětivějšího prostředí (tj. vývoj tzv. „user friendly interface“) pro komunikaci mezi strojem a člověkem.

Mezi nezbytné součásti této „přátelskosti“ patří bezpochyby komunikace s počítačem přirozeným jazykem a ne jen pomocí mechanických vstupních zařízení, jakými jsou myš, nebo klávesnice. Jednou z možností, jak umožnit komunikaci počítače a člověka pomocí mluvené řeči, je rozpoznávání hlásek a jejich konverze do podoby příkazů nebo psaného textu.

Samohlásky

Samohlásky jsou takzvané *znělé fonémy*, přičemž *foném* je nejmenší jednotka řeči, jež rozlišuje v daném jazyce slova. Fonémů bývá různé množství. Uvádí se, že v některých jazycích se vyskytuje až 60 různých fonémů. *Znělý foném* je takový foném, jenž zachovává periodičnost a kumuluje energii do určitých frekvencí – tzv. *formantů*. Formantů je několik a lze je jednoznačně definovat pomocí hodnoty frekvence, v níž je kumulováno nejvíce energie. Jelikož vznik formantů je vázán na dutiny hlasového traktu, jímž prochází řečový signál, lze i jednotlivé formanty přiřadit jednotlivým větším dutinám a poloha formantu pak reprezentuje rezonanční kmitočet dané dutiny, jež zde působí jako dutinový rezonátor a ovlivňuje tak velkou měrou signál.

Vlastností samohlásek v českém jazyce je možnost jejich jednoznačné klasifikace pomocí prvních dvou formantů. První formant F_1 odpovídá dutině hrdelní a druhý formant F_2 pak dutině ústní. Podobně formant F_3 odpovídá dutině nosní, ale ta nemá na českou mluvu takový vliv, jako první dvě uvedené. Pomocí formantů F_1 a F_2 lze tedy samohlásky klasifikovat podle tabulky 1.

Pakliže však existuje takový vztah mezi prvními dvěma formanty řečového signálu, pak je otázkou, proč je vlastně tak těžké rozpoznávat hlásky. Odpověď je poměrně jednoduchá. Každý z nás je individualita, každý mluví jiným způsobem, má jiné zlozvyky, jinou výslovnost, jiný hlasový trakt, jiný hlasový rozsah. Je tolik faktorů, v nichž se naše

hlasy liší, že se tímto úkol rozpoznání hlásek stává velice náročným. Rozdílnost těchto parametrů hlasu však není největší potíž. Problém činí především takzvané souzvuky. Hlásky vyslovené jedna za druhou v rychlém sledu, podle toho, jak rychle mluvíme, se vzájemně více, či méně ovlivňují. Toto vzájemné ovlivňování se velice projeví na výsledném charakteru řeči. Ovšem pravdou také je, že samohlásky jsou hlásky, které lze, i přes toto ovlivňování, velmi dobře rozpoznat.

Samohláska	F_1 [Hz]	F_2 [Hz]
A	800 – 1000	1200 – 1400
E	500 – 700	1600 – 2100
I	300 – 500	2100 – 2700
O	500 – 700	900 – 1200
U	300 – 500	600 – 1000

Tabulka 1.: Vztah českých samohlásek k formantům F_1 a F_2

Neuronová síť typu „Backpropagation“

Vícevrstvé neuronové sítě typu Backpropagation [2][3] se pro svoje vlastnosti využívají především jako klasifikátory, neboť mají výbornou schopnost oddělit od sebe určité, předem specifikované skupiny. Odezvou sítě je v tomto případě míra podobnosti předloženého neznámého vzoru a vzorů, na něž byla síť naučena. Toho může být využito například pro rozpoznání číslic, znaků a jiné druhy rozpoznávání. Další využití tohoto typu sítě spočívá v realizaci různých transformací vstupních proměnných. Obecně je tato síť schopna provést téměř libovolnou transformaci (při přiměřeně zvoleném rozsahu sítě). Takovými transformacemi mohou být například komprese signálů, kdy je v jedné ze skrytých vrstev (příp. ve výstupní vrstvě) snížen počet neuronů, čímž dochází k redukci (resp. kompresi) dat vstupního vektoru na určitým způsobem specifikovaný výstup. Další oblastí využití těchto sítí je filtrace signálů, neboť filtrace je určitý druh transformace, kterou lze realizovat právě pomocí neuronové sítě.

Neuronová síť typu RCE (Restricted Coulomb Energy)

Síť typu RCE [2] má velmi rychlý algoritmus učení a stačí pouze velice málo iteračních kroků, aby se tato síť naučila velice kvalitně reagovat na vstupní vektory. Úspěšnost správného zařazení vstupního vektoru, pokud je to vektor, který byl naučen, je 100%, což je vynikající. Problém však nastává u vstupních vektorů, na něž síť naučena nebyla. Takovéto vektory, zvlášť při velkém počtu prvků vstupu, mohou být velice často chybně klasifikovány. Čímž se využití této sítě poněkud omezuje a k rozpoznávání samohlásek se příliš nehodí, neboť zde je rozpětí hodnot vstupních vektorů tak veliké, že úspěšnost rozpoznávání samohlásek v plynulé řeči je velice malá (v porovnání např. se sítí typu Backpropagation).

Nicméně i přesto lze tuto síť využít v mnoha jiných aplikacích (především pro 100% úspěšnost rozpoznání naučených vzorů).

Kohonenova neuronová síť

Kohonenova [2] síť se také označuje jako samo se organizující mapa. Tato síť sama vybere výrazné rysy vstupních vektorů a rozdělí je do oblastí, které posléze odpovídají určitým třídám. Díky tomu, že jsou významné příznaky vektorů sítí automaticky rozpoznány, tj. síť se učí bez učitele, hodí se tato síť tam, kde lze této možnosti využít. Také se ukazuje, že tato síť má vcelku dobré vlastnosti právě v oblasti rozpoznávání řeči. Nicméně úkolu rozpoznávat samohlásky se nezhostila tak dobře, jako třeba síť typu Backpropagation.

LVQ (Learning Vector Quantization)

Metoda LVQ [3] je založena na dodatečném naučení sítě učící se bez učitele (např. Kohonenova síť) pod dohledem. Tj. síť naučená bez učitele je poté přeučena s učitelem, čímž se výrazně zvýší úspěšnost klasifikace. Experimentálně bylo dokázáno, že v případě analýzy řeči, se výsledky dávané sítí zlepšily o 50%, tj. úspěšnost rozpoznání byla o polovinu větší, neboť pomocí druhého stupně učení se eliminují případy chybné klasifikace, a tím se účinnost sítě zvýší.

Závěr

Vliv koartikulace a velká závislost úspěšnosti rozpoznání na mluvčím negativně ovlivňuje výsledky rozpoznávání, implementovaných sítí a parametrů použitých pro reprezentaci jednotlivých hlásek.

Z uvedených neuronových sítí se nejlépe hodí sítě typu Backpropagation a síť typu LVQ. Tyto projevovaly nejlepší vlastnosti v pokusech s cizími mluvčími a plynulým textem. Především pak síť typu BPN nebyla tak citlivá na nevhodně zvolené parametry. V tomto směru je síť typu LVQ poněkud náročnější, neboť nepatrná změna parametrů (např. počet kroků učení) způsobila nedoučení nebo naopak přeučení sítě. Mezi nevýhody této sítě pak také patří vcelku dlouhý čas učení při velkém počtu neuronů.

Každopádně samotná neuronová síť není vhodná pro náročnější, komplexnější aplikace, jako je např. diktovací psací stroj. V tomto případě je nutné pak dodatečné přepracování například s využitím skrytých Markovových řetězců, kterážto metoda se ve spojení s neuronovými sítěmi ukazuje jako nejlepší., případné využití jiné metody založené na pravděpodobnosti.

[1] *Rodman D.R.:* Computer Speech Technology, Boston, Mass.: Artech House, 1999

[2] *Zakharian S., Ladewig-Riebler P., Thoer S.*: Neuronale Netze für Ingenieure, Wiesbaden: Vieweg, 1998

[3] *Orság, F.*: Rozpoznávání samohlásek v plynulé řeči, Semestrální projekt, VUT FEI Brno, 2000