MODELLING PROSODIC DYNAMICS FOR SPEAKER RECOGNITION

Zdeněk Jančík
Master Degree Programme (2), FIT BUT
E-mail: xjanci03@stud.fit.vutbr.cz

Supervised by: Pavel Matějka
E-mail: matejkap@fit.vutbr.cz

ABSTRACT

The current automatic speaker recognition systems extract speaker-dependent features by looking at short-term information. This article focuses on long-term information about speakers. I used approach that use the fundamental frequency and energy trajectories for each speaker.

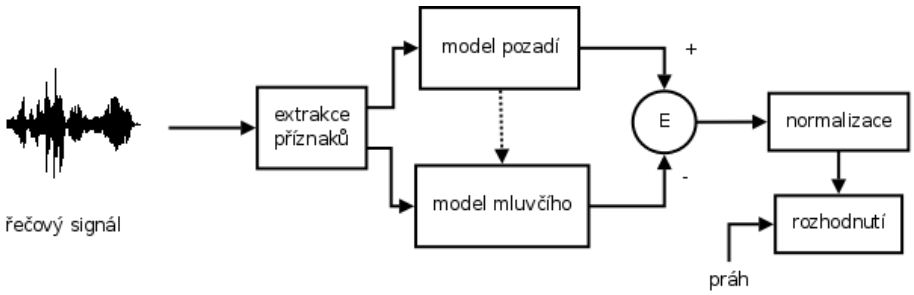
1 ÚVOD

Příznaky použitelné pro rozpoznání řečníka můžeme rozdělit na krátkodobé a dlouhodobé. Krátkodobé příznaky se získávají z řečových rámců dlouhých asi 20 až 30 milisekund. Naopak dlouhodobé příznaky se extrahují ze segmentů dlouhých asi stovky milisekund.
V článku se zaměřím na dlouhodobé příznaky založené na prosodii. Podle [1] patří mezi prosodické vlastnosti řeči výška hlasu, hlasitost a časování. Tyto vlastnosti můžeme modelovat trajektorií základního tónu a krátkodobé energie (podle [2]).
Systémy založené na prosodických příznacích sice ve srovnání s tradičními[1] mají nižší úspěšnost rozpoznání, ale protože využívají pro rozpoznání jiný druh informací, tak se s výhodou jejich výsledky kombinují s tradičními systémy, čímž lze dosáhnout relativního zlepšení až o desítky procent (podle [2]).

2 ROZPOZNÁNÍ MLUVČÍHO

Na obrázku 1 je blokové schéma platné pro většinu systémů pro rozpoznání mluvčího.
Při rozpoznání mluvčího si pro každého mluvčího natrénujeme model cílového mluvčího. Další speciální model UBM (Universal Background Model, model pozadí) natrénujeme na všech trénovacích datech. Výsledné skóre je rozdílem skóre modelu pro daného mluvčího a UBM.
Nejpoužívanější přístupy pro modelování jsou GMM, skryté Markovovy modely, neuronové sítě a SVM(Support vector machine).
Normalizací skóre myslíme upravení hodnot skóre do určitého rozsahu, případně normalizací počtem příznaků získaných ze vstupního souboru.

[1] používající krátkodobých příznaků jako jsou například MFCC koeficienty(Mel frekvenční kepstrální koeficienty [1]) modelované pomocí GMM (Gaussians mixture model [1])



Obrázek 1: Schéma systému pro rozpoznání mluvčího

V praxi se pro zvýšení úspěšnosti často spojují výsledky více systémů založených na různých příznacích.

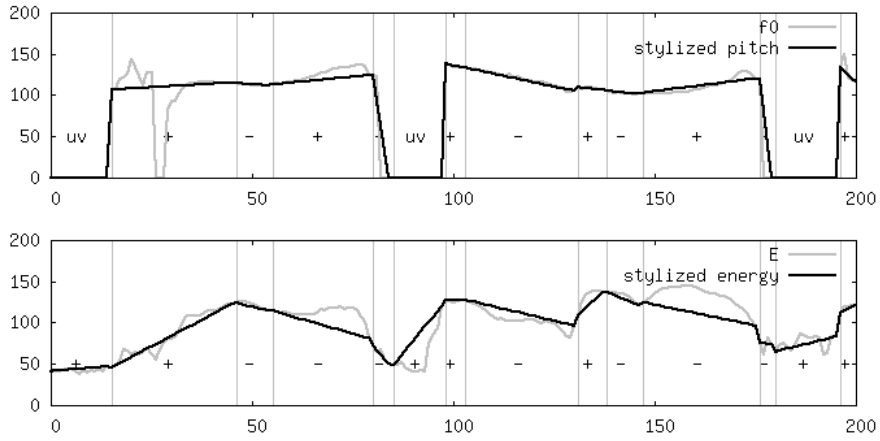
3 SEGMENTACE PODLE ZÁKLADNÍHO TÓNU

Pomocí programu *praat* (`http://www.fon.hum.uva.nl/praat/`) rozdělíme vstupní signál na rámce o délce 20 ms a přesahu 10 ms. Pro každý rámec určíme frekvenci základního tónu a energii.

Detekce základního tónu v programu praat je založena na normalizované autokorelaci. Z kandidátů na frekvenci základního tónu je ten nejlepší vybrán viterbiho algoritmem tak, aby šlo o nejlepší cestu prostorem kandidátů. Podrobný popis algoritmu je v [3].

Pokud si průběh základního tónu a energie vyneseme do grafu (obrázek 2), jasně odlišíme segmenty, kde základní tón stoupá, klesá, případně je nulový (jde o neznělý rámec). Řečový signál potom rozdělíme na takové segmenty.

Pokud veličina stoupá, označíme segment značkou +, pokud klesá −. Ze základního tónu získáme první značku a z energie druhou (pokud je segment neznělý označíme jej `uv`). Tuto dvojici můžeme dále rozšířit o délku segmentu (dlouhý, krátký, střední) a foném odpovídající segmentu.



Obrázek 2: Výsledky segmentace

Před vlastní segmentací je třeba vyhladit průběh základního tónu mediánovým filtrem. Tímto omezíme chyby detekce základního tónu a vyhladíme průběh základního tónu, což povede ke vzniku menšího počtu delších segmentů.

Pokud bychom při segmentaci nijak neomezili minimální délku segmentu vzniklo by velké množství krátkých segmentů a rozpoznání by nebylo příliš úspěšné. Proto segmentaci řeším ve dvou krocích. Nejprve vygeneruji všechny segmenty, které se potom spojuji tak, abych eliminoval segmenty kratší než určitý práh (3 až 5 rámců). Přesnou délku prahu je nutné experimentálně ověřit.

4 TRÉNOVÁNÍ A TESTOVÁNÍ

Pro modelování mluvčích jsem použil bigramový a trigramový jazykový model. Trénování a testování je implementováno jako sada skriptů v programovacím jazyce `bash`, volajících programy `ngram-count` a `ngram` z balíku SRILM (více informací o SRILM je na webu `http://www.speech.sri.com/projects/srilm/`).
Pro trénování a testování jsou použita NIST SRE 2006 data.

5 VÝSLEDKY A ZÁVĚR

	Systém	Chybovost (ERR)
1	bigram (f0,E)	35,92 %
2	trigram (f0,E)	37,15 %
3	bigram (f0,E) jiná segmentace	35,38 %
4	Andre na NIST SRE 2001 datech	20,3 %

Systémy 1 a 2 se liší použitým jazykovým modelem. Systém 3 používá mediánový filtr zachovávající svislé hrany a upravený algoritmus pro spojování segmentů. Pro srovnání je uveden i systém 4 ze článku [2] ovšem testovaný na jiných datech.
Výsledky dosažené samostatným systémem založeným na prosodických příznacích jsou obecně horší než výsledky nejlepších systémů založených na MFCC (zdroj [4]). Výsledky jsou ovšem vypočteny z jiných vlastností řečového signálů. Proto se systémy založené na modelování dynamiky prosodie kombinují s tradičními systémy, což vede ke značnému zlepšení výsledků.
Další práce na projektu budou zahrnovat přidání délky segmentu, využití výstupu fonémového rozpoznávače a kombinaci s tradičním systémem založeným na MFCC.

REFERENCE

[1] Psutka J. a kol.: Mluvíme s počítačem česky, Praha, Academia 2006.

[2] Adami Andre G. et al.: Modeling Prosodic Dynamics for Speaker Recognition, ICASSP 2003.

[3] Paul Boersma (1993): Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proceedings of the Institute of Phonetic Sciences 17: 97-110. University of Amsterdam, 1993.

[4] webová stránka SuperSID: `http://www.clsp.jhu.edu/ws2002/groups/supersid/`