TOWARDS RECOGNIZING PARALINGUISTIC EXPRESSIONS FROM EVERYDAY DIALOGS

Jan Mašek
Master Degree Programme (2), FEEC BUT
E-mail: xmasek09@stud.feec.vutbr.cz

Supervised by: Hicham Atassi
E-mail: atassi@feec.vutbr.cz

ABSTRACT

This document describes three methods for the classification of paralinguistic expressions such as laughing and crying from spoken dialogs. An original database with three states was built for this purpose. When analyzing human's everyday dialogs, especially in films and TV series, music might appear, so the built database was extended by three new states, namely: music, singing with music and neutral speech with background music. This gave us six classes for the classification. Feature extraction, feature reduction, classification and fusion are common steps in recognizing for all three methods presented. The difference between the methods consists in the classification approach. The first method, called *straight approach*, classifies all classes considered at once. The second method called *decision tree oriented approach*, exploits five sub-classifiers in the tree structure. The last method uses for classification *emotion coupling approach* proposed in [1]. The best features were identified by using a statistical method called *F*-ratio and GMM classifiers were used in all methods under examination.

1. ÚVOD

Paralingvistika zkoumá součásti řeči, které se nedají vyjádřit písemným projevem. Může se jednat například o rozdíly ve výšce hlasu, tempu, artikulaci, intenzitě nebo o smích, pláč, vzdychy a sténání. Často se stává, že v případě rozpoznání emocí se některé paralingvistické stavy chybně klasifikují například jako vztek. Při analýze videonahrávek se často setkáváme i s hudbou. V této práci se proto budeme zabývat mimo jiné i klasifikací smíchu, pláče a hudby. V mnoha ohledech lze klasifikaci těchto stavů srovnávat s klasifikací emocí, kterými se zabývali např. K. Soltani a R. N. Ainon [2] nebo H. Atassi [1].

2. ROZBOR

Základem procesu rozpoznání vzorů je rozsáhlá databáze nahrávek, která je vhodně rozdělena podle hledaných tříd. Druhým a nejdůležitějším krokem je výpočet příznaků. Může se jednat o příznaky v časové, spektrální a kepstrální oblasti, nebo o příznaky transkripční. Snahou je vypočítat co nejvíce příznaků, které dokážou nejlépe oddělit dané třídy. Dalším krokem je redukce příznaků a posledním úkolem je jejich klasifikace pomocí vhodného klasifikátoru.

2.1. DATABÁZE NAHRÁVEK

Úkolem bylo vytvořit obsáhlou databázi nahrávek pro rozpoznání nejen paralingvistických signálů v řeči. Jedná se o nahrávky smíchu, pláče, neutrální promluvy, hudby, hudby se zpěvem a řeči s hudbou v pozadí. Všechny nahrávky byly uloženy ve formě souborů wav se vzorkovacím kmitočtem 16kHz, rozlišovací schopností 16bitů a jedním mono kanálem. V databázi je celkem 651 nahrávek od 83 mluvčích, které zastupují šest hledaných stavů. Nahrávky obsahující neutrální promluvu byly pořízeny od mluvčích české národnosti.

2.2. VÝPOČET A REDUKCE PŘÍZNAKŮ

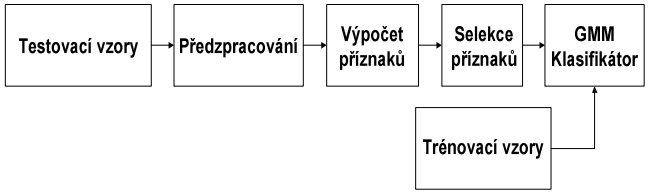
Pro každý experiment bylo vypočteno 621 segmentálních a 12768 suprasegmentálních příznaků. Přehled segmentálních příznaků lze nalézt v tab. 2.1.

Redukce probíhala pro každou klasifikaci zvlášť za pomocí metody F-poměru. Příznaky jsou seřazeny dle kvality a část z nich je pak použita při klasifikaci.

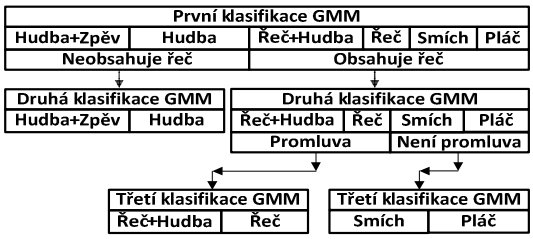
20 Mel-frekvenčních kepstrálních koeficientů+Δ+ΔΔ, 18 Koeficientů percepční lineární predikce+Δ+ΔΔ, 20 Koeficientů spektra melovské banky filtrů+Δ+ΔΔ, 30 Koeficientů kepstra, 18 Koeficientů lineární predikce, Energie, Teagrův operátor energie, Počet průchodů nulou	*Délka segmentu*: 512
	Překrytí: 256
Spektrální příznaky	*Délka segmentu*: 1024
	Překrytí: 512
Mel-spektrální modulační energie na 4Hz, Mel-kepstrální modulační energie na 4Hz	*Délka segmentu*: 640
	Překrytí: 320

Tab. 2.1: Přehled segmentálních příznaků

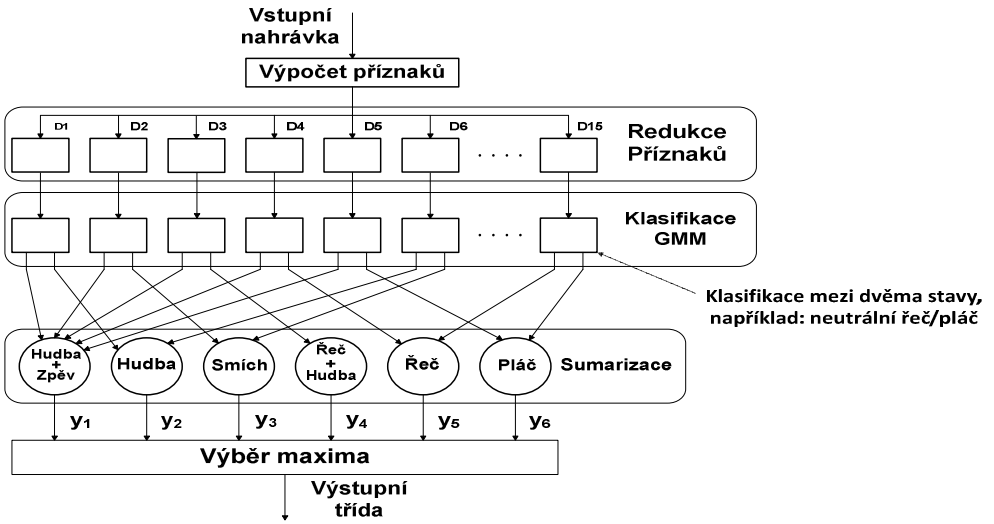
2.3. NAVRŽENÉ EXPERIMENTY



Obr. 2.1: Postup klasifikace při prvním experimentu



Obr. 2.2: Postup klasifikace při druhém experimentu



Obr. 2.3: Postup klasifikace při třetím experimentu

	mi	mu	la	ms	ne	cr
mi	63	21	12	3	1	0
mu	26	72	1	0	1	1
la	4	3	78	1	4	9
ms	2	0	0	78	20	0
ne	0	0	1	24	75	0
cr	0	0	4	0	0	96
Celková úspěšnost [%]						77,0

Tab. 2.2: Matice záměn pro první experiment

	mi	mu	la	ms	ne	cr
mi	66	30	0	4	0	0
mu	8	92	0	0	0	0
la	2	8	62	4	22	2
ms	8	4	0	80	8	0
ne	0	0	2	2	96	0
cr	0	2	6	0	0	92
Celková úspěšnost [%]						81,3

Tab. 2.3: Matice záměn pro druhý experiment

	mi	mu	la	ms	ne	cr
mi	62	30	6	2	0	0
mu	8	92	0	0	0	0
la	6	6	78	0	8	2
ms	6	0	2	84	8	0
ne	2	0	2	2	94	0
cr	0	0	4	0	0	96
Celková úspěšnost [%]						84,3

Tab. 2.4: Matice záměn pro třetí experiment

Legenda: **mi**-hudba se zpěvem, **mu**-hudba, **la**-smích, **ms**-řeč s hudbou v pozadí, **ne**-neutrální řeč, **cr**-pláč

3. ZÁVĚR

Z dosažených výsledků prvního experimentu vyplývá, že klasifikace všech tříd najednou není příliš úspěšná. Počet 500 použitých příznaků je vysoký a průměrná úspěšnost klasifikace při plné kovarianční matici GMM klasifikátoru se nachází pod 80%. Klasifikátor zde nejlépe detekoval pláč a nejhůře hudbu se zpěvem. Nejvíce chyb vznikalo v rozpoznávání mezi hudbou/hudbou se zpěvem a mezi neutrální řečí/neutrální řečí s hudbou v pozadí.

Druhý experiment v podstatě zjednodušuje úlohu klasifikace zavedením více podtříd, které lze mezi sebou snáze odlišit. Kvůli tomu se zvýšil počet klasifikací na 5, tím se ale snížil počet potřebných příznaků (průměrně na 72 pro každý klasifikátor). Výpočetní nároky klasifikátoru se dále snížili použitím diagonální kovarianční matice. Celková úspěšnost klasifikace zde činí 81,3% a nejhůře se zde oddělovaly podtřídy hudba se zpěvem/hudba a smích/neutrální řeč.

Poslední experiment se snažil rozvést myšlenku z druhého experimentu ještě dál a to klasifikací všech dvojic samostatně. Dílčí výsledky klasifikace naznačují velký potenciál této metody jako univerzálního řešení pro podobné úkoly. Celková úspěšnost klasifikace se pohybuje nad 84%. Nejhůře se opět rozlišují třídy hudba se zpěvem/hudba a smích/neutrální řeč.

LITERATURA

[1] ATASSI, H., RIVIELLO, M., SMÉKAL, Z., HUSSAIN, A., ESPOSITO, A.: Emotional Vocal Expressions Recognition using the COST 2102 Italian Database of Emotional Speech. Lecture Notes in Computer Science (IF 0,513), 2009, č. 5967, s. 1-14. ISSN: 0302- 9743.

[2] SOLTANI, K., AINON, R.N.: Speech emotion detection based on neural networks, Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on , vol., no., pp.1-3, 12-15 Feb. 2007

[3] DUDA, R., HART, P., STORK, D.: Pattern Classification. Wiley, 2003