DISTRIBUTIVE SPEECH RECOGNITION FOR EMBEDDED SYSTEMS

Jan KRYŠTOF, Master Degree Programme (5)
Dept. of Computer Systems, FIT, BUT
E-mail: xkryst04@stud.fit.vutbr.cz

Supervised by Dr. Lukáš Burget

ABSTRACT

The aim of the project is to develop a system for distributive speech recognition for embedded systems. The target devices are low-end cellular phones. The system consists of two parts. The client-side application situated on the cellular phone will perform the extraction of speech features, particularly Mel Frequency Cepstral Coefficients (MFCC) which will be sent to the server-side application. This application accepts the speech features and uses it as an input for the external speech recognizer. The recognizer uses the Viterbi algorithm for the speech recognition. After recognition an answer with the result is sent from the server to the client.

1 ÚVOD

V posledních letech enormně vzrostl zájem o mobilní telefony a multimediální aplikace, které se stávají běžným programovým vybavením. Multimedialní aplikace pracují obvykle s velkými objemy dat a nárokují si velké množství procesorového času. Výpočetní výkon mobilních telefonů obvykle nenabízí možnosti pro běh aplikací jako stolní počítače a proto jsou možnosti aplikací u mobilních telefonů limitovány. Jednou z takovýchto aplikací je i rozponávání řeči, například rozpoznávání lizolovaných slov u hlasového vytáčení telefonních čísel v adresáři.

Proces rozpoznávání řeči můžeme rozdělit do třech fází: získání dat reprezentujících řeč, extrakce řečových příznaků, a rozpoznání nad řečovými příznaky. U současných mobilních telefonů probíhají tyto tři fáze v rámci jednoho programu, běžícím v telefonu. Tento projekt přichází s myšlenkou rozdělit proces rozpoznávání mezi telefon a mezi jinou, výpočetně silnější jednotku. Jinak řečeno, nechat mobilní telefon provést operace, na které jeho výkon stačí, a zbylou část zadat stroji s větším výpočetním výkonem. Jedná se tedy o systém klient – server, kde v roli klienta vystupuje mobilní telefon a v roli serveru počítač.

Cílem projektu je vytvořit knihovnu pro klientskou aplikaci pro mobilní telefony, které disponují platformou Java. Knihovna bude obsahovat funkce pro extrakci řečových příznaků a funkce umožňující přenos dat mezi klientem a serverem. Dále pak implementovat server s funkcí rozpoznávače a navrhnout komunikační protokol pro výměnu dat mezi klientem a

serverem.

2 KLIENTSKÁ APLIKACE

Klientská aplikace přijímá hlasový vstup od uživatele a provádí extrakci řečových příznaků, konkrétně MFCC (Mel Frequency Cepstral Coefficients). Klient je napsaný pro platformu J2ME s profilem MIDP 2.0 (The Mobile Information Device Profile) a CLDC 1.0 (Connected Limited Device Configuration). Konfigurace definuje programové vybavení (zahrnuje Virtual machine a API). V současnosti jsou na trhu mobilních telefonů v užívání dvě konfigurace: 1.0 a 1.1. CLDC 1.0 byla vybrána vzhledem k velikosti uživatelské základny, takže koncová aplikace nediskriminuje vlastníky starších telefonů s Javou a CLDC 1.1 je zpětně kompaktibilní s 1.0.

Aplikace přistupuje k audio systému mobilního telefonu a na pokyn uživatele snímá vzorky signálu řeči. Třídy a metody pro přístup k audio systému jsou poskytované volitelným balíčkem MMAPI 1.1 (Mobile Media API). Hlasový signál je vzorkován na 8 kHz, a je na něj aplikována procedura provádějící extrakci příznaků, konkrétně MFCC (Mel Frequency Cepstral Coefficients) .

2.1 EXTRAKCE ŘEČOVÝCH PŘÍZNAKŮ

Hlasový signál je segmentován na rámce o délce 200 vzorků. Start každého dalšího rámce je vzdálen 80 vzorků od předchozího rámce, takže překrýv je 120 vzorků. Nad takto získanými rámci jsou prováděny následující operace.

Obr. 1: *Proces extrakce řečových příznaků (MFCC)*

Váhování Hammingovým oknem. Hodnotám signálu je udělena váha pomocí funkce Hammingova okna o délce 200. Takto váhovaný signál je doplněn nulami (zero - padding), takže výsledný signál má delku 256.

Fourierova transformace. Signál je dále transformován do frekvenční oblasti pomocí diskrétní Fourierovy transformace (FFT), konkrétně pomocí rychlé FFT, která redukuje složitost výpočtu $\mathbf{O}(N^2)$ -> $\mathbf{O}(N \log N)$. Výsledkem operace je spektrum signálu, v tomto případě 256 komplexně sdružených čísel. Díky symetrii spektra bereme pouze polovinu hodnot (129), ze kterých jsou spočítány $|F_{k|}$ koeficienety spektra.

Váhování MEL bankou. Koeficienty spektra jsou dále váhovány sadou trojúhelníkových oken – MEL bankou. Banka obsahuje celkem 23 oken, což má za následek redukci vstupu o 129 hodnotách na 23 hodnot. Z hodnot získaných z MEL banky je dále spočítán logaritmus.

Diskrétní kosinová transformace. Poslední krok spočívá ve výpočtu 23-bodové Diskrétní kosinové tranformace (DCT) nad výstupem z MEL banky. Tranformace je promítnuta do prvních třinácti bází, takže výstupem je 13 hodnot, 13 MFCC.

Je vidět, že po výpočtu MFCC došlo k velké redukci hodnot signálu (80:13), což vede ke zrychlení komunikace mezi serverem a klientem.

2.2 FIXED POINT ARITMETIKA

U implementace knihovny funkcí pro výpočet řečových příznaků je třeba zohlednit omezení, které přináší CLDC 1.0. Tato konfigurace nepodporuje datový typ čísel s plovoucí desetinnou čárkou. Tento problém je vyřešen použitím fixed-point aritmetiky, což znamená, že reálná čísla jsou simulována pomocí celých čísel. Podstata věci je tato: datový typ integer má 32 bitů, z nichž lze 16 bitů použít pro informaci o celém čísle před desetinnou čárkou a zbylých 16 bitů nese informaci o desetinné části. Jedná se tedy o fixed-point 16::16. Použití fixed-point aritmetiky je dokonce v tomto případě výhodou, jelikož aritmetické operace s celými čísly jsou rychlejší než s reálnými a snahou je zrychlit výpočet co nejvíce.

3 SERVER

Aplikace představující server je napsaná pro platformu Java 2 SE. Je určená IP adresou, naslouchá na určitém portu a odpovídá na požadavky klienta. Server je schopný volat rozpoznávač, který byl dodán již na startu projektu. Klient obdrží od serveru jako výsledek informaci o výsledku rozpoznávání.

Rozpoznávač pracuje na principu Viterbiho algoritmu využívající skryté Markovovy modely.

Rozpoznávač potřebuje pro svoji funkci databázi modelů získanou natrénováním nad hlasovými daty. Databázi lze získat dvěmi způsoby. Použít již natrénovanou databázi na nějakém původním systému a zjistit jakým způsobem je nutné upravit příznaky z klientské aplikace tak, aby byla použitelná pro původní databázi. Druhá možnost spočívá v natrénovaní vlastní databáze, tak, že telefon bude přijímat řečová data, provádět extrakci příznaků a data odesílat na server do databáze.

ZÁVĚR

Projekt je zaměřen na low-end mobilní telefony a jeho cílem je prozkoumat možnosti distribuovaného rozpoznávání řeči a zhodnotit výpočetní výkon mobilních telefonů a rychlost bezdrátového přenosu dat u těchto zařízení. Projekt je v tomto čase ve fázi, kdy je klient schopen zasílat řečová data od uživatele a posílat je na server. Byly provedeny testy výkonu procesorů mobilních telefonů a provedena analýza možností vývoje multimediálních aplikací pro mobilní telefony, pro platformu Java 2 ME.Podrobné informace o tomto projektu si lze přečíst v diplomové práci “**The Distributive Speech Recognition for Embedded Systems**”, která vzniká souběžně s tímto článkem.

REFERENCES/LITERATURA

[1] Psutka, J.: Komunikace s počítačem mluvenou řečí, ACADEMIA, Praha 1995

[2] Kinnunen, T., Karpov, E., Fränti, P.: Automatic Speaker Recognition for Series 60 Mobile Devices, University of Joensuu, Finland 2004

[3] Jan, J.: Číslicová filtrace, analýza a restaurace signálu,Vutium, Brno 2002