ZMĚNA DÉLKY TRVÁNÍ SYNTETIZOVANÉ ŘEČI METODOU WSOLA

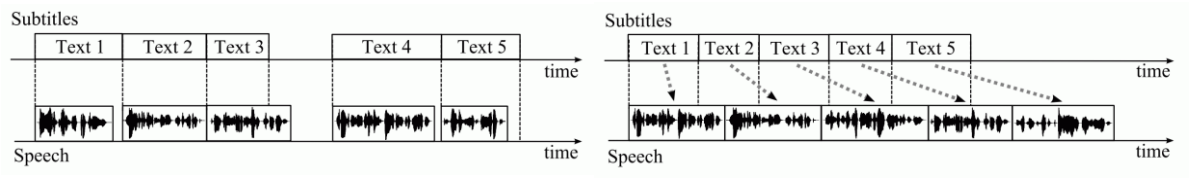
Martin Méner[1]

1 ÚVOD

Tato práce se zabývá možnostmi změny délky trvání syntetizované řeči metodou WSOLA (Waveform Similarity Over-Lap Add). Hlavní důraz je kladen na využití této metody v projektu ELJABR (Eliminace jazykových bariér handicapovaných diváků České televize), jehož cílem je zpřístupnit vysílání České televize širší skupině diváků. Celý projekt je možné rozdělit na dvě hlavní části, a to syntézu a rozpoznávání řeči. Této práce se využije právě v části zabývající se syntézou, jejímž obecným úkolem je automatické vytváření řečového signálu využitím existujících titulků jednotlivých pořadů. K tomu je použit TTS (Text-to-Speech) systém ARTIC, jehož vstupem jsou titulky běžně dostupné prostřednictvím teletextu a výstupem je syntetizovaná zvuková stopa. Vzhledem k motivaci tohoto projektu je nová zvuková stopa syntetizována tak, aby byla srozumitelnější, pomalejší a zcela bez rušivých zvuků na pozadí. V této práci jsou nejprve diskutovány důvody, kvůli kterým je v některých případech nutné změnit délku trvání syntetizované řeči a poté je presentována metoda WSOLA, která se nyní k modifikacím využívá a jsou zmíněny výsledky této práce a návrhy dalších vylepšení do budoucna.

2 DŮVODY PRO ZMĚNU DÉLKY TRVÁNÍ

Výstupem obecného systému TTS je nejlepší možná syntetizovaná zvuková stopa bez ohledu na délku, neboť přesná délka syntetizované řeči obvykle není vyžadována. Proto vzhledem ke specifičnosti použití může dojít k časovému posunu mezi televizním pořadem a novou zvukovou stopou. Pokud je doba promluvy kratší nebo stejně dlouhá jako časový interval příslušného titulku, popřípadě když řeč přesahuje interval svého titulku, ale stále ještě nezasahuje do časového intervalu titulku následujícího, pak je přes popsanou desynchronizaci celý dialog pochopitelný. Pokud ovšem promluva překračuje nejen příslušný interval titulku, ale i interval následující, dojde ke zpožďování promluvy vůči obrazu, a tak se může dialog stát nepochopitelným. Oba případy jsou ukázány na obr. 1.

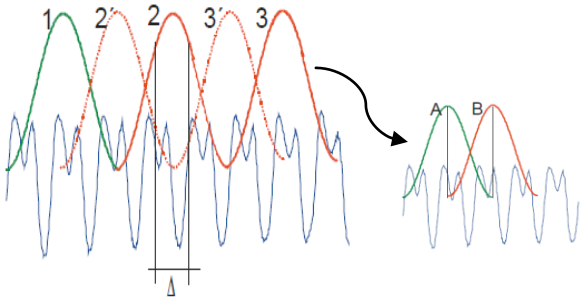


Obr. 1: Časová desynchronizace titulků

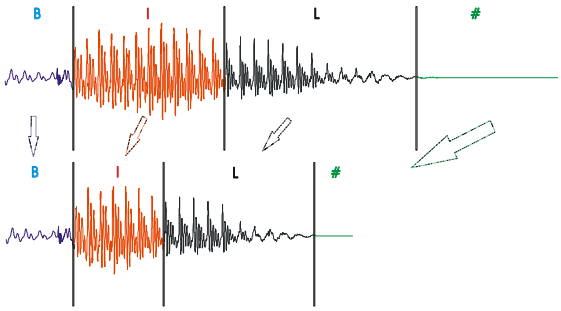
[1] Bc. Martin Méner, student navazujícího studijního programu Aplikované vědy a informatika, obor Kybernetika a řídicí technika, e-mail: mmener@students.zcu.cz

3 METODA WSOLA

Metoda WSOLA rozvíjí metodu OLA (Over-Lap and Add) a v projektu ELJABR se využije ke zkracování promluvy, aby nedocházelo k výše popsaným problémům v dialozích. Princip metody spočívá v dělení vstupního signálu na menší segmenty a následně v jejich přeskládání, k čemuž se využívá von-Hannovo okénko. Přeskládání je dáno modifikací, kterou má algoritmus vykonat. Na rozdíl od metody OLA je ale nyní při přeskládání využita podobnost jednotlivých segmentů tak, aby charakter výstupu odpovídal vstupu. Jako kritérium optimality je využita metoda nejmenších čtverců. Průběh algoritmu zkrácení fonému je na obr. 2. První vstupní segment (pozice 1) je přidán na výstup (pozice A). Poté začne hledání druhého segmentu (pozice 2), který bude přidán na pozici B. Poloha druhého segmentu je dána jednak zkrácením, které je požadováno, a jednak podobností se vzorem, který je na pozici 2‘. Algoritmus pak dále pokračuje přidáním třetího segmentu, který bude mít nyní vzor na pozici 3‘ a „nejpodobnější“ segment se bude hledat v okolí pozice 3. Algoritmus je dále rozšířen o rozlišení mezi jednotlivými fonémy, kdy v současné době se rozlišují pauzy, samohlásky, souhlásky a explozivy. Algoritmus se nejprve pokusí zkrátit pauzy, pokud zkrácení nestačí, modifikují se samohlásky a popřípadě nakonec souhlásky. Příklad modifikace délky v závislosti na jednotlivých fonémech je ilustrován na obr. 3.



Obr. 2: Algoritmus WSOLA



Obr. 3: Zkrácení slova „byl“ [bil], # je pauza

Algoritmus přináší velmi dobré výsledky pro zkracování výstupního signálu až na 70% signálu vstupního. V budoucnu je možné zlepšit určování koeficientů zkrácení citlivějším dělením fonémů a to i s přihlédnutím ke statistickým datům z řečového korpusu tak, aby modifikace řeči metodou WSOLA více korespondovala se způsobem zkracování řeči člověkem. Dále je možno zapracovat na využití metody k prodlužování řeči.

Poděkování: Příspěvek byl podpořen grantovým projektem SGS-2010-054.

LITERATURA

Demol M., Verhelst W., Struyve K., Verhoeve P., “Efficient Non-Uniform Time-Scaling of Speech with WSOLA". *Proceedings of the Speech and Computers 2005 (SPECOM–2005)*, pp. 163–166, Patras, Greece, 2005.

Hanzlíček Z., Matoušek J., Tihelka D., “Towards Automatic Track Generation for Czech TV Broadcasting: Initial Experiments with Subtitles-to-Speech Synthesis“. *Proceedings of the 9th International Conference on Signal Processing, ICSP'08*, vol. 3, pp. 2721-2724, IEEE Press, Beijing, China, 2008.

Méner M., Tihelka D., “The Possibilities of Time Scale Modification of Speech*”. Proceedings of the 19th Czech-German Workshop on Speech Processing*, pp. 107-113, Prague, Czech Republic, 2009.