ПРИЛАГАНЕ НА ТЕХНОЛОГИЯТА НА HTML5 ЗА ПРЕДСТАВЯНЕ НА АУДИО И ТРАНСКРИПЦИЯ НА РАЗГОВОРНА РЕЧ

Хетил Ро Хауге
(Осло, Норвегия)

HTML5, нов стандарт за уеб страници, предлага нови възможности за представяне на транскрибирана устна реч: потребителят може да премине директно от произволна реплика в транскрипцията към съответното място в аудио файла. Изработването на подобни страници е сравнително трудоемко, но може да се прави с безплатни програмни продукти.

HTML5 е най-новият вариант на HTML, езикът за описание на уеб страници и други форматирани документи (W3C 2011). Този стандарт на уеб страници предоставя нови възможности за интегриране на текст с аудио и видео и следователно за представяне на елементи от устната комуникация. Това се прави с помощта на прости процедури за предаване на аудио и видео – за сравнение при предишните стандарти са необходими сравнително сложни структури при програмирането на страницата, както и инсталиране на plugin като QuickTime, Flash, Silverlight или др. от страна на потребителя.

Стандартът е все още в процес на подготовка и не всички браузъри го поддържат: Internet Explorer само след версия 9, Firefox – след версия 7, Google Chrome – след версия 15 и Safari – след версия 5.

Кои приложения на тази технология могат да бъдат интересни за изследователите на устната комуникация? HTML5 дава възможност за създаване на уеб страници, които комбинират текст с аудио или видео с хиперлинкове директно от даден ред в текста към съответното място в аудиото или видеото. В илюстрираната от фиг. 1

демонстрационна програма (Хауге 2011а), направена за проекта BGSpeech (изследователска група за изследване на устната комуникация, Софийски университет), кликване върху дадено изречение в дясната колона пренасочва плейъра отляво към съответното място в аудиото (едновременно същото изречение се изписва като субтитри към плейъра).

[Фиг. 1, bgspeech.jpeg, тук]

Тази функционалност се извършва с помощта на Javascript библиотеки, които се намират на уеб сървъра. Плейърът е jPlayer (http://jplayer.org/) и връзката между пълния текст и видеото със субтитрите се осъществява от popcorn.js (http://popcornjs.org/). Демонстрационната програма, направена за BGSpeech, е минимален модел – възможно е да се направят още рамки на уеб страницата с текст, който се променя динамично, следвайки аудиото или видеото, например с научен коментар.

За добри резултати е необходимо да се направи ръчно кодиране на съответствията между текста, от една страна, и аудиото или видеото, от друга страна. Записът трябва да се прослуша в реално време и при всяко ново изречение да се регистрира времевата точка на сегмента (обикновено с кликване на дадено копче в програмния продукт). Удобен програмен продукт за тази цел е EXMARaLDA (Extensible Markup Language for Discourse Annotation, безплатна програма, http://www.exmaralda.org/). EXMARaLDA предлага експортиране в XML - разпространен стандарт за структуриране на текст. Съществуват безплатни редактори за XML за най-разпространените компютърни платформи, например EditX Lite (http://free.editix.com/).

В представената по-долу част от файл от демонстрационната програма за BGSpeech стойностите T2, T3, ... са разстоянията в секунди от дадено начало, при които се сменят говорни "събития" (events). При самите събития, т.е. транскрипцията, във втората част от файла, са отбелязани началната и крайната точка като препращания към списъка на T-стойности.

```
<tli id="T2" time="5.986665105894844"/>
<tli id="T3" time="13.839995407747294"/>
<tli id="T4" time="15.839995407747294"/>
```

[...]

<event start="T2" end="T3">изпит (.) и че няма верни и грешни отговори още повече че (.) специално за тематиката която ше говориме днеска ъ_ъ_ъ </event>

<event start="T3" end="T4">(.) със сигурност вие сте по-компетентни </event>

С помощта на така наречени XSL-трансформации (опция в редактор за XML) XML-структури могат да бъдат конвертирани в: чист текст, HTML, други XML-структури, или javascript. В посоченотата демонстрационна програма дисплеят с пълния текст се записва в HTML и "събитие" от горния пример изглежда така:

<p>Ясен [v] : изпит (.) и че няма верни и грешни отговори още повече че (.) специално за тематиката която ше говориме днеска ъ_ъ_ъ

За целите на представяне на отделните изречения в плейъра същото изречение е записано в javascript по следния начин:

```
.subtitle({
  start: 5.986665105894844,
end: 13.839995407747294,
text:"Ясен [v]: изпит (.) и че няма верни и грешни отговори още повече че (.) <b>специално</b> за тематиката която ше говориме днеска ъ_ъ_ъ ",
```

Интересна идея е, че успоредяването на текст и аудио може да се прави автоматично, ако условно се приеме, че всяко изречение заема толкова голям относителен дял в аудиото, колкото в писмената транскрипция. От само себе си се разбира, че по такъв начин няма да се получи точността на ръчното кодиране. Но дали точността ще е достатъчна за ориентир, ако не за представителни цели пред публика, то поне за изследователски цели, когато се обработва аудио в по-голям размер, например около час или повече, с транскрипция, а без кодиране?

Донякъде е предимство, ако графичната система на даден език отразява една фонема само с една графема – както е горе-долу в български, руски, чешки, фински, за разлика например от полски,

английски и скандинавските езици. По-добре е аудиото да бъде "равномерно" – не дискусия или разговор, а в оптималния случай – спокоен монолог. Като предварителна подготовка за такъв процес аудиото трябва да бъде обработено, като се отстранят паузите в двата края, както и елементи, които не присъстват в текста. И обратно, в текста трябва да бъдат отстранени елементи, които не се срещат в аудиото.

Текстът се сегментира на изречения, например с общодостъпната програма за разделяне на изречения на проекта за паралелен германско-романско-славянски корпус при Университета в Осло (Orekhov б.г). След това "формалното" успоредяване на транскрипцията на час или два часа запис се изчислява за секунди с AppleScript на платформата Macintosh. Подобни резултати могат да се постигнат с Perl на други платформи.

Най-добрият постигнат резултат е с повестта на Франц Кафка, In der Strafkolonie (В наказателната колония, Хауге 2011б). Аудиото, записано от доброволец, приносител на librivox.org, е 69 минути и 30 секунди (броячът на плейъра jPlayer започва броенето отново от всеки час и показва 09.30 минути на посочената страница) и попаденията обикновено са +/- едно до две, понякога три изречения.

В по-реалистичен пример от устната комуникация, актуално интервю, 2 минути и 45 секунди, в сутрешния блок на полското радио, (Хауге 2011в), попаденията стават по-неточни към края на записа. Участниците обаче говорят много бързо и оживено и резултатите може би биха били по-добри при по-формална или по-спокойна устна комуникация.

БИБЛИОГРАФИЯ

Хауге 2011а: Hauge, Kjetil Rå. Фокус група.// http://folk.uio.no/kjetilrh/bgspeech/fokus-grupa.html.

Хауге 2011б: Hauge, Kjetil Rå. Automatisk parallellstilling av tekst og audio.// http://folk.uio.no/kjetilrh/wkshop/kafka/inderstrafkolonie.html.

Хауге 2011в: Hauge, Kjetil Rå. Polsk intervju.// http://folk.uio.no/kjetilrh/pol/popcorn/ intervju.html.

Orekhov б.г.: Orekhov, Boris. RuN Project Tools Page.// http://nevmenandr.net/run/tools/.

W3C 2011: HTML5: A vocabulary and associated APIs for HTML and XHTML. W3C Working Draft 25 May 2011.// http://www.w3.org/TR/2011/WD-html5-20110525.

APPLYING HTML5 TECHNOLOGY FOR DISPLAYING AUDIO AND TRANSCRIBED SPOKEN LANGUAGE

Kjetil Rå Hauge
(Oslo, Norway)

Abstract: HTML5, a new web standard, opens new possibilities for displaying transcribed spoken language: the user may jump directly from any turn in the transcript to the appropriate place in the audio. Preparing pages for this purpose is relatively time-consuming, but may be done with freely available computer programs.