NAKI-II-UJC-UKONCENE - Task #4161

Automatický přepis MONO dat

09.02.2017 15:17 - Zajíc Zbyněk

Status:	Closed	**Start date:**	09.02.2017
Priority:	Normal	**Due date:**	
Assignee:	Psutka Josef V.	**% Done:**	70%
Category:		**Estimated time:**	0.00 hour
Target version:			

Description

Až bude nový LM, vytvořit přepis dat.

History

#1 - 10.04.2017 08:07 - Zajíc Zbyněk

- Assignee changed from Pražák Aleš to Psutka Josef V.

- % Done changed from 0 to 70

První výsledky na mono datech - na ~27k slovech ~3h reci je to 8kHz mono desiva kvalita.

LM (slovník)	Corr[%]	Acc[%]
1.2M	51.49	44.86
174k	67.33	60.17

AM:
The first experiment was made with the low-quality data without the distinguished channels (both the language counselor and the client of LCC stored in one channel). We applied our recent triphone HMM acoustic model. The basic speech unit was a three-state HMM with 32 mixtures of multivariate Gaussians for each of the 4969 states. The model was trained on various 500~hours of spontaneous telephone speech, all converted into low quality (8kHz, μ-law resolution). We used the PLP parameterization as our front-end module (19 band pass filters, 12 cepstral coefficients with delta and delta-delta features with CMN).

LM:
1.2M
Our initial ASR system is using the universal trigram back-off Language Model (LM) with the mixed-case vocabularies with more than 1.2M words. Our training text corpus contains the data from newspapers (520 million tokens), web news (350 million tokens), subtitles (200 million tokens) and transcriptions of some TV programs (175 million tokens).

174k
For the better aim of the language model, we trained a domain LM with the dictionary size 174k as a standard trigram language model with Kneser-Ney smoothing. This model was trained on the available transcribed data from LCC of the language counselor (220 thousand tokens) and the client (180 thousand tokens) and from the email communication (counselor 3,8 million tokens and client 3,4 million tokens).

#2 - 18.11.2019 09:01 - Zajíc Zbyněk

- Status changed from Assigned to Closed