Some Issues affecting the Transcription of Hungarian Broadcast Audio

Anindya Roy[1], *Lori Lamel*[1], *Thiago Fraga da Silva*[1], *Jean-Luc Gauvain*[1], *Ilya Oparin*[2]

[1]Spoken Language Processing Group, CNRS-LIMSI, Orsay, France.
[2]Laboratoire National de métrologie et d'Essais (LNE), Paris, France.
{roy, lamel, thfraga, gauvain}@limsi.fr, ilya.oparin@lne.fr

10.21437/Interspeech.2013-676

Abstract

This paper reports on a speech-to-text (STT) transcription system for Hungarian broadcast audio developed for the 2012 Quaero evaluations. For this evaluation, no manually transcribed audio data were provided for model training, however a small amount of development data were provided to assess system performance. As a consequence, the acoustic models were developed in an unsupervised manner, with the only supervision provided indirectly by the language model. The language models were trained on texts downloaded from various websites, also without any speech transcripts. This contrasts with other STT systems for Hungarian broadcast audio which use at least 10 to 50 hours of manually transcribed data for acoustic training, and typically include speech transcripts in the language models. Based on mixed results previously reported applying morph-based approaches to agglutinative languages such as Hungarian, word-based language models were used. The initial Word Error Rate (WER) of the system using context-independent seed models from other languages of 59.8% on the 3h development corpus was reduced to 25.0% after successive training iterations and system refinement. The same system obtained a WER of 23.3% on the independent Quaero 2012 evaluation corpus (a mix of broadcast news and broadcast conversation data). These results compare well with previously reported systems on similar data. Various issues affecting system performance are discussed, such as amount of training data, the acoustic features and choice of text sources for language model training.

Index Terms: Large vocabulary continuous speech recognition (LVCSR), broadcast news transcription, Hungarian language, unsupervised training, agglutinative languages, Bottleneck MLP features

1. Introduction

With 17 million native speakers, Hungarian is the most widely spoken *non*-Indo-European language in Europe, spoken mostly in Hungary, but also in Austria, Croatia, Romania, Serbia, Slovakia, Slovenia and Ukraine. Hungarian is highly agglutinative and inflected [1]. Each verb may have 50 prefixed forms (on average), 59 inflections and many verb-to-verb derivations. Each noun may have about 900 inflections. This leads to large lexica, high out-of-vocabulary (OOV) rates and data sparsity for language modeling, making automatic speech recognition quite challenging [2]. However, it has a close to phonemic orthography, simplifying the creation of pronunciation dictionaries.

Although several Hungarian speech-to-text (STT) systems have been reported previously [2][3][4][5][6][7][8][9], the language still remains relatively less investigated. In particular, the authors know of only two other major works dealing specifically with STT for Hungarian *broadcast* audio [10][11]. Both systems used manually transcribed speech for acoustic model training.

In contrast, this paper describes the development of an STT system for Hungarian broadcast audio which used *unsupervised* training of acoustic models [12][13][14] in accordance with one of the Quaero project goals of low-cost system development. Manual transcriptions were available only for 3 hours of development data which were used to tune the coefficients to interpolate language models trained from different text sources and assessing system versions. Furthermore, no extensive language-specific knowledge was applied. No native Hungarian speakers were involved. It was decided to use words as lexical units instead of morphs [2][11] because of mixed results from previous studies which compared morph-based and word-based approaches. In fact, for Hungarian broadcast audio, it was shown that a word-based recognizer can outperform morph-based ones if sufficient amount of textual data is available for training language models [10].

This paper discusses the effect of various issues on WER, including choice of suitable text corpora for language modeling and vocabulary selection. Importantly, it was found that appending probabistic features computed using bottleneck Multilayer Perceptron (MLP) [28] to standard PLP+F0 features can reduce the WER by about 6.5% (absolute) and 16% (relative) on average, even when the MLP is trained using speech from an unrelated language. To the best of the authors' knowledge, this is the first time that such a finding has been made for the Hungarian STT task. The paper is organized as follows: Section 2 describes the transcription system. Section 3 details the experiments carried out, the optimal system configuration and evaluation results. Section 4 concludes the work.

2. System description

The system is based on the LIMSI broadcast news transcription system and has two components: the audio partitioner and word recognizer. An overview of the partitioner is provided here since it does not form an original contribution of this work (see [15] for details): A maximum likelihood segmentation/clustering iterative procedure is applied to the audio, segmenting it first into speech and non-speech segments. The speech segments are then clustered into individual speakers. Once the audio has been partitioned, each speaker cluster is processed by the word recognizer. The following paragraphs outline the components of the Hungarian word recognizer (language and acoustic models).

2.1. Language models

The system uses n-gram language models. Language model development involved collection and preprocessing of suitable

Text source	# sentences	# words (total)	# words (unique)
Google News (GNews)	4.88M	83.1M	1.67M
Wikipedia (HWiki)	3.86M	56.0M	2.02M
Hunglish Corpus (HC)	1.58M	11.2M	565K
Quaero dev12 corpus	947	26.8K	8.41K

Table 1: Text corpora after normalization used for LM training. Quaero dev12 was used to tune LM interpolation coefficients (ref. Sec.2.1).

text sources, vocabulary creation, training LMs from each text source and interpolating them.

2.1.1. Collection of text sources

The following text corpora were located, downloaded and processed to use for language modeling: (1) *GNews* All articles in Hungarian from Google News[1] aggregated from 2009 to 2012. (2) *HWiki* The complete Wikipedia in Hungarian[2], and (3) *HC* The Hunglish Corpus [16], which has three sections, (a) modern literature, (b) classical literature, and (c) movie subtitles.

Multiple iterations of cleaning and normalization were performed on these corpora [17][18], with particular attention to the following aspects: First, there were many sentences in English (particularly in GNews) which had to be removed. This was difficult to do automatically, as many Hungarian words are spelled exactly in the same way as an English word with different meaning, e.g. *fog, hold, nap, mint, most*, etc. Finally, the chosen solution was to remove any sentence with a sequence of 3 or more consecutive English words. Second, numbers were converted into words, taking into account spoken forms for dates, time, money, etc. In Hungarian, numbers are agglutinated, i.e. 137 would be written as *onehundredthirtyseven* and not *one hundred thirty seven* as in English. This complicates the task since each number is a new word. In this work, number were segmented into constituent parts and each part was considered as a word. Third, the first letter of words were uncapitalized where required (for example, common nouns). For this, the first word was removed from each sentence and the ratio of number of times a word appeared in the resulting corpora with its initial letter capitalized to the number of times with the initial letter small was calculated. Words with this ratio more than or equal to 5% were left untouched, the rest were uncapitalized. This reduced the case-sensitive WER of the system. In general, it took more effort to clean/normalize the raw text from HWiki than from GNews or HC. Relevant statistics of the cleaned and normalized text corpora are detailed in Table 1.

2.1.2. Vocabulary creation

It can be seen in Table 1 that total number of unique words in the text corpora is very high. Many of these words may not appear in broadcast news transcripts. Hence, a suitable vocabulary was created as follows: First, a full word list was created from all text sources. Next, unigram models were trained on each text source individually using this word list and interpolated, with the mixture weights computed via EM algorithm to minimize perplexity on manual transcriptions of the Quaero 2012 development (dev12) corpus [19] (ref. Section 3.1) consisting of 3 hours of Hungarian broadcast audio. The final vocabulary was created by selecting words with a unigram LM probability higher than a preset threshold, chosen according to size of the vocabulary desired. Different vocabulary sizes were considered: 100K, 200K, 500K and 1M. Words were directly used in the vocabulary. No lexical decomposition steps were performed.

2.1.3. Training & interpolating LMs

Once the vocabulary was created, 2-gram, 3-gram and 4-gram LMs were trained on each text source individually using modified Kneser-Ney smoothing. For GNews, each year was treated as a separate text source. Each set of LMs (2-,3-,4-gram) were then interpolated using the transcriptions of the Quaero dev12 corpus [19].

2.2. Acoustic models

The system uses triphone-based left-to-right context-dependent HMMs [20][21], with tied-states. Each state output is modeled by a mixture of 32 gaussians, except for silence which is modeled using 96 or 1024 gaussians. Steps involved in acoustic model development are creation of phone set, creation of pronunciation dictionary, and unsupervised training of acoustic modes.

2.2.1. Phone set

An initial set of 40 distinct phones was chosen, in addition to silence, filler words and breath sounds. However, it was found that training data was insufficient to model some rarely-occuring phones which were discarded and replaced by other similar phone(s). For example, the phone /dz/ was later replaced by the sequence /d//z/. Although not strictly equivalent, this change improved system performance. The long forms of vowels were treated as separate phones. However, separate phones were not created for the long forms of consonants (gemination). They were taken into account in the pronunciation dictionary (ref. Section 2.2.2). The final list contained 37 phones (13 vowels, 24 consonants).

2.2.2. Pronunciation dictionary

Since Hungarian has a close to phonemic orthography, a set of grapheme-to-phoneme (g2p) rules was initially constructed [22]. These rules were evaluated on a separate pronunciation dictionary created from the online Wiktionary database[3] of about 6K most frequent Hungarian words. Based on feedback from these evaluations, a few modifications/improvements on the g2p rules were made in several iterations including: (1) a few instances of consonant assimilation, such as /t/ $\sim$ /d/, /k/ $\sim$ /g/, /m/ $\sim$ /n/, and (2) inclusion of glide /y/ between 2 adjacent vowels, when one of them is /i/, such as /i/ /a/ $\rightarrow$ /i/ /y/ /a/. The final set of g2p rules performed accurately (with more than 95% correct match) on the Wiktionary database. Two solutions were proposed for dealing with repeated consonants (which translate to long/geminated forms of the consonants): (1) treating long consonants as doubled instances of the short forms of the same consonants, (2) replacing long consonants by a single instance of the short form (i.e. ignoring gemination), creating two dictionaries: *Dict1* and *Dict2* respectively. A third solution was initially proposed using additional phones for these geminates (as had been investigated before for Arabic [23] and Italian),

[1] http://news.google.com/ The authors would like to thank Vocapia Research for help with downloading the articles.

[2] http://dumps.wikimedia.org/backup-index.html

[3] http://en.wiktionary.org/

Phone symbol	Seed model language	Phone description
a	English	As in *call*
A	English	As in *cat*
d	French	As in French *début*
u	French	As in *rude*
ö	German	Similar to *i* in *bird*
j	Italian	As in *jam*
ç	Russian	As in *check*
ñ	Russian	As in Spanish *niña*

Table 2: Seed models from different languages (ref. Sec.2.2.3).

Language	Vowels	Consonants	Total
English	7	15	22
French	4	2	6
German	1	0	1
Italian	1	1	2
Russian	0	6	6
Total	13	24	37

Table 3: Distribution of seed models from different languages.

but this were discarded in preference to a smaller phone set. The pronunciation dictionary was created by obtaining pronunciations for each word in the vocabulary using the established g2p rules. A seperate set of g2p rules were applied to acronyms based on how each letter is named, rather than their phonetic value (e.g., in English, *m* → /e/ /m/ rather than *m* → /m/, etc).

2.2.3. Unsupervised training of acoustic models

For training acoustic models, a corpus of 370 hours of unlabeled audio broadcast by MR1 Kossuth Rádió[4] and InfoRádió[5] channels in 2011 was used. The AMs were trained in an unsupervised way [12][13][24][14]. First, seed models for each phone were chosen from pre-trained models for English, French, German, Italian and Russian [25][26]. The choice of seed model for each phone were made by the authors by (1) noting their IPA equivalent cross-checked using various sources (such as [22], http://en.wikipedia.org/wiki/Hungarian_alphabet), and (2) listening to examples of Hungarian words and matching the sound with their existing knowledge of sounds in the other languages. Table 2 lists a subset of phones used and the language from which their seed models were chosen while Table 3 shows the distribution of the seed models among the five languages. It is observed that most of the models are from English, French or Russian with a few from Italian and German. These seed models were first used to decode a small subset of the training audio (75h) and the system hypotheses were used as groundtruth reference transcripts for re-training the models using ML. A few iterations of decoding and re-training models were carried out in the same way, gradually increasing the amount of raw audio data decoded from 75h to 370h and consequently number of contextual phones modeled from 735 to 20882. This iterative process led to successively more accurate models. Initially, PLP+F0 features were used [27]. After some training iterations, when a reasonably robust model was ready, probabilistic Bottleneck MLP features [28] were appended to the PLP+F0 features.

Layer	No. of units
Input layer	475
Hidden layer 1	3500
Hidden layer 2 (MLP features)	39
Output layer	108

Table 4: Topology of MLP used to generate MLP features.

2.2.4. Appending Bottleneck MLP features

A four-layer MLP with a narrow third layer in the middle (bottle-neck) was used to compute the MLP features. The topology of the MLP is provided in Table 4. The MLP uses modified TRAP-DCT features as input [28]. The output of the second hidden layer (39-dimensional) is taken as the MLP feature set. The MLP was trained on 95 hours of manually transcribed audio data in *English* (a mix of broadcast news and conversations collected previously under the Quaero project), with automatic phone state segmentations [30]. The MLP used phone-state targets during training.

3. Experimental evaluation

3.1. Corpora and methodology

Two speech corpora were used for the experiments: (1) Quaero 2012 development (dev12) corpus, and (2) Quaero 2012 evaluation (eval12) corpus. These corpora were collected and transcribed as part of the 2012 Quaero STT evaluation campaign. Each set contains about 190 minutes of audio broadcast by MR1 Kossuth Rádió and InfoRádió channels during September-November 2011 and include both read news (70% of the total duration) and spontaneous conversations (30% of the total duration). Segments were recorded both on-site and in studio with on-site segments often noisy. There were 192 speakers in total, with 972 speaker turns. There is no overlap of data in the training, dev12 and eval12 corpora.

The system was first evaluated on the dev12 corpus to study the effect of various issues on WER. Based on these studies, the optimal system configuration was chosen. The decoding parameters (LM scaling factor, penalties for word and silence) were tuned on the dev12 corpus. Once configured and tuned, the resulting system was used to process the eval12 corpus as part of the Quaero 2012 benchmark, with results reported by an external entity, the LNE,[6] running the evaluation. Results were reported in terms of both case-insensitive (CI) and case-sensitive (CS) WERs.

3.2. Study of different issues

Duration of audio for acoustic training The first three rows of Table 5 shows how WER on the dev12 corpus reduces gradually from 59.8% to 40% as the duration of audio used for acoustic model training and number of contextual phones modeled were increased with each training iteration (ref. Section 2.2.3). The third training iteration used about 370 hours of audio. PLP+F0 features were used. The vocabulary size was 200K words, GNews corpus was used for LM training.

Appending BN MLP features The last row in Table 5 shows that appending MLP features to PLP+F0 after the third training iteration led to significant reduction of WER (6-7% absolute and 16% relative) although the MLP was trained using English

[4] http://www.mr1-kossuth.hu/

[5] inforadio.hu

[6] Laboratoire National de métrologie et d'Essais (www.lne.fr)

Iter-ation#	Duration of audio for AM training	# contextual phones modeled	WER (CI)	WER (CS)
1	75h	735	59.8	61.7
2	91h	3983	55.0	57.3
3	370h	20882	40.0	43.4
4	370h + **MLP**	20882	**33.0**	**37.2**

Table 5: Effect of duration of training audio on WER (%), on Quaero dev12 corpus. First 3 rows: PLP+F0 features used. Last row shows reduction in WER on addition of MLP features after third training iteration. Voc. size = 200K words. GNews corpus used for LM training.

Vocab. size	OOV	PPX	WER (CI)	WER (CS)
100K	7.9	569.4	34.3	35.5
200K	5.4	771.4	31.6	32.8
500K	3.2	1030.5	30.1	31.3
1M	2.4	1176.3	29.5	30.7

Table 6: Effect of vocabulary size on OOV (%), PPX (4g) & WER (%) on Quaero dev12 corpus. GNews corpus used for LM training.

speech. This shows the advantage of such features in a cross-lingual setting [29].

Vocabulary size Table 6 shows the the effect of vocabulary size (ref. Section 2.1.2) of LM in terms of out-of-vocabulary (OOV) rate, PPX (4-gram LM) and WER on Quaero dev12 corpus. The WER drops by about 1.7%, going from 100K to 200K, 1.5% from 200K to 500K and 0.6% from 500K to 1M. The GNews corpus was used for LM training. At this stage, further normalization of LM training texts were performed, leading to further WER reduction (compare row 4 in Table 5 to row 2 in Table 6, both used GNews corpus and 200K words for LM training).

Choice of corpora Table 7 summarizes experiments aiming to assess the relevance of the available text corpora for the transcription of Hungarian broadcast audio. The perplexity (PPX) for 4-gram LM and WER using the Quaero dev12 corpus are given for various setups: LMs trained on individual sources and their interpolation in pairs or using all 3. The vocabulary size was 1M words. Individually, GNews performed significantly better than others, while HC performed the worst. The dominant performance of GNews could be explained in terms of (1) matching domain (news), (2) larger amount of text (83.1M words, ref. Table1) and (3) cleaner text. Interpolating LMs (either in pairs or all 3) only slightly reduced WER over the best component LM.

Miscallaneous issues A second decoding pass with MLLR/ CMLLR adaptation reduced the WER by 1.8%, as shown in Table 8. Using gender-dependent AMs reduced the WER by 0.5% compared to gender-independent AMs. A silence model using a mixture of 96 Gaussians reduced the WER by 0.2% compared to one using 1024 Gaussians. For modeling geminates, both dictionaries Dict1 and Dict2 were tried (ref. Sec. 2.2.2), the former performing slightly better.

3.3. Evaluation results

Based on the above studies, an optimal system was created using component LMs with 1M word vocabulary trained on GNews, HWiki and HC and interpolated on dev12, dictionary Dict1, gender-dependent AMs using PLP+F0+MLP trained on 370 hours of audio with 20882 contextual phones, silence model using 96 Gaussians and MLLR/ CMLLR adaptation. This system achieved a WER of 27.7% (CI) and 28.9% (CS) on dev12 and 25.7% (CI) and 26.9% (CS) on the eval12 corpus. After the Quaero evaluations, the models were retrained once more, using existing models to decode the 370h of training data. This resulted in a further reduction of the WER to 25.0% (CI) and 26.3% (CS) on the 3 hour dev12 corpus, and 23.3% (CI) and 24.6% (CS) on the 3 hour eval12 corpus.

Corpus	PPX	WER (CI)	WER (CS)
HC	6603.0	37.8	39.8
HWiki	2543.7	32.6	37.1
GNews	1176.3	29.5	30.7
HWiki+HC	2119.0	31.0	36.2
GNews+HC	1155.0	29.4	30.7
GNews+HWiki	1113.8	29.3	30.6
GNews+HWiki+HC	1101.3	29.2	30.5

Table 7: Effect of choice of text corpora for LM training on PPX (4g) and WER (%) on Quaero dev12 corpus. '+' denotes interpolation of component LMs. Voc. size = 1M words.

	WER (CI)	WER (CS)
1st pass	29.5	30.7
2nd pass	27.7	28.9

Table 8: Reduction of WER (%) with second decoding pass, using Quaero dev12 corpus. PLP+F0+MLP features used, voc. size = 1M words. GNews corpus used for LM training.

The current system compares reasonably well with previous studies on Hungarian broadcast audio. For example, the best system reported in [10] used 50 hours of transcribed speech for acoustic model and language model training and achieved a WER of 26.3% (CI) on 49 minutes of broadcast news and 49.4% (CI) on 52 minutes of broadcast conversations. Another system [11] which used 10 hours of transcribed speech for training achieved a WER of 21.0% on 1 hour of broadcast news. Note that the current system was trained in an unsupervised way. Also, the evaluation data is longer (3 hours) and is a mix of broadcast news (70%) and conversations (30%).

4. Conclusion

This work presents a transcription system for Hungarian broadcast audio. Using words as lexical units and unsupervised acoustic training, it achieved a WER of about 24% on 2 independent 3 hour sets of development and evaluation data, comparing well with existing systems which used supervised acoustic training. Appending MLP features significantly improved performance over PLP+F0 features, even though the MLP was trained using English data.

5. Acknowledgements

This work has been partly financed by OSEO, the French State Agency for Innovation, under the Quaero program. Research by the fifth author was done while he was with CNRS-LIMSI.

6. References

[1] L. Tihanyi, "Number of Hungarian Word Forms," *Multilingual Text Tools and Corpora for Central and Eastern European Languages Project COPERNICUS 106, Deliverable D1.2. Language-specific resources - Appendix 2*, 1996, http://aune.lpl.univ-aix.fr/projects/multext-east/MTE2.number.html.

[2] P. Mihajlik, *Recognition of Spontaneous Hungarian Speech without Language Specific Rules*, Ph.D. thesis, Budapest University of Technology and Economics, Budapest, Hungary, 2010.

[3] M. Szarvas, *Efficient large vocabulary continuous speech recognition using weighted finite-state transducers - The development of a Hungarian dictation system*, Ph.D. thesis, TITECH, Tokyo, Japan, 2003.

[4] Gy. Zsigri, L. Toth, A. Kocsor, and Gy Sejtes, "Az automata s kzi szegmentls ejtsvaricik okozta problmai," in *Proceedings of Magyar Szamitogepes Nyelveszeti Konferencia*, 2004.

[5] K. Vicsi and Gy. Szaszak, "Examination of pronunciation variation from hand-labelled corpora," in *Proceedings of International Conference on Text, Speech and Dialogue*, 2004.

[6] G. Szaszak, *Szupraszegmentlis jellemzk szerepe s felhasznlsa a beszdfelismersben*, Ph.D. thesis, Budapest University of Technology and Economics, 2008.

[7] A. Banhalmi, D. Paczolay, L. Toth, and A. Kocsor, "Investigating the robustness of a hungarian medical dictation system under various conditions," *International Journal of Speech Technology*, vol. 9, no. 3-4, pp. 121–131, 2008.

[8] P. Mihajlik, Z. Tuske, B. Tarjan, B. Nemeth, and T. Fegyo, "Improved recognition of spontaneous hungarian speech - morphological and acoustic modeling techniques for a less resourced task," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1588–1600, 2010.

[9] G. Sarosi, T. Fegyo, P. Mihajlik, B. Tarjan, J. Pancza, and Z. Hans, "LVCSR-based Speech Analytics of a Hungarian Language Call-Center," in *IAST 2012: Workshop on Innovation and Applications in Speech Technology*, 2012.

[10] B. Tarjan, P. Mihajlik, A. Balog, and T. Fegyo, "Evaluation of lexical models for hungarian broadcast speech transcription and spoken term detection," in *Proceedings of CogInfoCom 2011: 2nd International Conference on Cognitive Infocommunications*, 2011.

[11] P. Mihajlik and B. Tarjan, "On Morph-based LVCSR Improvements," in *Proceedings of SLTU*, 2010, pp. 10–16.

[12] L. Lamel, J.-L. Gauvain, and G. Adda, "Unsupervised acoustic model training," in *Proceedings of ICASSP*, 2002.

[13] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: recent experiments," in *Proceedings of EuroSpeech*, 1999.

[14] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly Supervised Acoustic Model Training," in *Automatic Speech Recognition - Challenges for the New Millenium*, 2000, pp. 150–154.

[15] J.-L. Gauvain, L. Lamel, and G. Adda, "Partitioning and transcription of broadcast news data," in *Proceedings of ICSLP*, 1998, pp. 1335–1338.

[16] D. Varga, L. Nemeth, P. Halacsy, A. Kornai, V. Tron, and V. Nagy, "Parallel corpora for medium density languages," in *Proceedings of the RANLP*, 2005, pp. 590–596.

[17] G. Adda, M. Adda-Decker, J.L. Gauvain, and L. Lamel, "Text normalization and speech recognition in french," in *Proceedings of Eurospeech*, 1997.

[18] M. Adda-Decker, G. Adda, and L. Lamel, "Investigating text normalization and pronunciation variants for german broadcast transcription," in *Proceedings of Interspeech*, 2000, pp. 266–269.

[19] L. Lamel and B. Vieru, "Development of a speech-to-text transcription system for finnish," in *Proceedings of Spoken Language Processing and Understandin (SLTU)*, 2010.

[20] J.-L. Gauvain and C.-H. Lee, "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[21] M. Gales and S. Young, "The Application of Hidden Markov Models in Speech Recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2007.

[22] S. Grimes, "On the creation of a pronunciation dictionary for Hungarian," in *Proceedings of Midwest Computational Linguistics Colloquium, Urbana, Illinois*, 2006.

[23] A. Messaoudi, J.-L. Gauvain, and L. Lamel, "Arabic Broadcast News Transcription Using a One Million Word Vocalized Vocabulary," in *Proceedings of ICASSP*, 2006.

[24] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, no. 1, pp. 115–129, 2002.

[25] J. Loof, C. Gollan, and H. Ney, "Cross-language Bootstrapping for Unsupervised Acoustic Model Training: Rapid Development of a Polish Speech Recognition System," in *Proceedings of Interspeech*, Brighton, 2009.

[26] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised Cross-Lingual Knowledge Transfer in DNN-Based LVCSR," in *Proceedings of SLT*, Miami, 2012.

[27] H. Hermansky, "Perceptual Linear Prediction (PLP) Analysis for Speech," *Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, 1990.

[28] P. Fousek, L. Lamel, and J.-L. Gauvain, "Transcribing Broadcast Data Using MLP Features," in *Proceedings of Interspeech*, 2008.

[29] F. Grezl, M. Karafiat, and M. Janda, "Study of probabilistic and bottleneck features in multilingual environment," in *Proceedings of IEEE ASRU*, Hawaii, 2011.

[30] T. Fraga da Silva, V.-B. Le, L. Lamel, and J-L. Gauvain, "Incorporating MLP features in the unsupervised training process," in *Proceedings of SLTU 2012 Third International Workshop on Spoken Languages Technologies for Under-resourced Languages*, pages 30-34, Cape Town, South Africa, 2012.