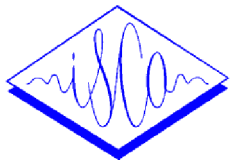
ISCA Archive
http://www.isca-speech.org/archive

SLTU-2014, St. Petersburg, Russia, 14-16 May 2014



SPEECH-TO-TEXT DEVELOPMENT FOR SLOVAK, A LOW-RESOURCED LANGUAGE

Cong-Thanh Do, Lori Lamel and Jean-Luc Gauvain

LIMSI-CNRS, Spoken Language Processing Group
B.P. 133, 91403 Orsay Cedex, France
E-mail: {ctdo, lamel, gauvain}@limsi.fr

ABSTRACT

Development of an automatic speech recognition (ASR) system for low-resourced languages is an important research topic in ASR. This paper reports on the development of a speech-to-text (STT) system targeting broadcast news and broadcast conversation transcription for the low-resourced Slovak language. Context-dependent acoustic models are trained without any manually transcribed audio data via cross-language transfer and unsupervised training. In addition, a pronunciation dictionary for Slovak language is created using efficient rule-based pronunciation modeling. For language modeling, large N-gram language models were estimated on 63M words of texts downloaded from the Internet. The system uses MLP (multi-layer perceptron) features imported from English which are concatenated with cepstral PLP (perceptual linear prediction) and F0 (pitch) features. These techniques were applied to develop a Slovak STT system with performance similar to that obtained by state-of-the-art systems for other languages. Furthermore, we propose to reduce the dimension of the MLP+PLP+F0 features from 81 to 50, using principal component analysis (PCA), in order to reduce the redundancy between the MLP and the PLP+F0 features. This feature reduction makes it possible to reduce the word error rate (WER) and the recognition time while reducing the CMLLR adaptation time by a factor of 3.

Index Terms— Slovak speech-to-text, ASR for low-resourced languages, Multi-layer perceptron, Unsupervised acoustic model training, Principal component analysis

1. INTRODUCTION

Development of an automatic speech recognition (ASR) system for a low-resourced language is one of the important research and development topics in ASR. This topic is the focus of recent research projects in speech processing, for instance the Quaero[1] and the Babel[2] projects. ASR technology has been initially developed for full-resourced languages, for instance English, French or Mandarin. Indeed, besides a large amount of text data needed for training language models, the development of ASR systems requires also acoustic data along with their transcriptions for training acoustic models. This is, in fact, the main technical point which limits the (rapid) development of an ASR system for a new low-resourced language since sufficient acoustic data with (manual) transcription is, often, not readily available for a new low-resourced language. Whilst large quantities of audio and text data can be downloaded from the Internet, for instance from radio and television news broadcast, the transcriptions of the audio data for training acoustic models are usually not available for low-resourced languages. Unsupervised acoustic models training [1] would be one of the solutions for training acoustic models without using transcriptions. Further, if large enough quantities of audio data can be found, the automatic transcriptions can be used to train language models [2] in case of representative text for language modeling of some under-resourced languages is difficult to obtain.

The Quaero project aims at developing ASR, or speech-to-text (STT), systems for European languages, including low-resourced languages, for instance Latvian [3], Hungarian [4] or Slovak. In this paper, we report the development of a state-of-the-art STT system for Slovak language, in the framework of the Quaero project. We are interested in the transcription of broadcast news (BN) and broadcast conversation (BC) data. The automatic transcription of BN and BC is a challenging task and has been studied for several years for full-resourced languages, such as English [5] or French [6]. Slovak language is a West-Slavic branch of European language which is spoken by 7 million people, mostly in Slovakia, and also in other countries, such as the Czech Republic or Hungary. BN and BC transcription systems for Slovak language have been developed in the literature [7]. However, more or less transcriptions of acoustic data have been utilized in the training of these systems.

2. CONTRIBUTIONS

2.1. System development

Our Slovak STT system development makes use of unsupervised acoustic models training [1], i.e., no transcriptions, and cross-language transfer acoustic modeling [8]. Cross-language transfer acoustic modeling has been applied in order to create context-independent (CI) acoustic models. The purpose of the cross-language transfer is to find common sound unit representations that are shared across languages. Indeed, the acoustic realization of phones could be similar for different languages since they are created through a limited set of articulatory movements [9]. Therefore, CI acoustic models of a low-resourced language can be initially selected from a set of CI acoustic models of available languages. These models are used as seed acoustic models in our STT system.

2.2. System improvement

Discriminative features, extracted by a trained multi-layer perceptron (MLP), have been introduced [10] and gradually adopted in state-of-the-art STT systems thanks to their relevance and effectiveness [11, 12, 13]. The implementation of MLP features in the STT systems at LIMSI makes it possible to improve significantly the recognition performance [11]. Generally, MLP features are used to augment cepstral features to create an augmented feature vector

[1] www.quaero.org
[2] http://www.iarpa.gov/Programs/ia/Babel/babel.html

[10, 11, 13]. The dimension of the MLP features is rather similar to that of the cepstral features. Hence, the dimension of the augmented feature vector is double that of the cepstral feature vector. This dimension doubling doubles also the numbers of parameters in the acoustic models using MLP features compared to acoustic models using cepstral features. Therefore, larger amounts of acoustic data are needed to estimate, reliably, the parameters of the acoustic models using MLP features.

In the development of low-resourced STT systems, speech and text data sparseness is a critical issue since the reliable estimation of the parameters of the statistical models needs a large amount of training data [14]. In this paper, besides the development of a baseline STT system for the low-resourced Slovak language, we are interested in the problem of speech data sparseness for training acoustic models using MLP features. As mentioned previously, with the doubling of the parameters of the acoustic models due to the use of MLP features, a larger amount of speech data is needed for training acoustic models. In the context of STT system development for low-resourced languages, namely Slovak, this issue should be taken into account.

The fact that augmenting cepstral features with MLP features improves the ASR performance demonstrates that the information conveyed in the MLP features is complementary to that conveyed in the cepstral features. However, the MLP and cepstral features could contain also redundant information to each other. It has been shown that reducing the dimension of the augmented feature vectors, using principal component analysis (PCA), helps in improving the *speaker verification* performance, compared to when using cepstral features alone [15]. Indeed, the PCA maintains the complementary but reduces the redundancy between MLP and cepstral features. Furthermore, the acoustic models will be more compact and less parameters are needed to be estimated if the dimension of the augmented feature vectors is reduced.

In the context of STT system development for low-resourced languages, this dimension reduction would be useful when lesser amounts of speech data is available for acoustic models training. Further, reducing the dimension of the feature vectors might help in reducing the recognition and adaptation time which is also an essential factor in the development of STT system. Indeed, when the dimension of the feature vectors is reduced, the transformation matrices used in the adaptation, for instance using MLLR (maximum likelihood linear regression) [16] or CMLLR (constrained MLLR) [17], are reduced, and hence, the adaptation time should be reduced. In this paper, we propose, thus, to apply the PCA to reduce the dimension of the augmented feature vectors which are used in our Slovak STT system.

The paper is organized as follows. Sections 3, 4, 5, and 6 present the development of the STT system, including the collection of text and acoustic data for language and acoustic models training, the building of N-gram language models, the pronunciation modeling, the acoustic features extraction and acoustic models training, respectively. After that, the experimental results are introduced and discussed in section 7. Finally, section 8 concludes the paper.

3. DATA COLLECTION FOR SYSTEM DEVELOPMENT

Speech and text resources are needed for training acoustic and language models. Our work aims at building a STT system to transcribe BN and BC data. In our work, broadcast speech data are downloaded from Slovak radio sources, available through their websites. The 3 radio sources are Slovensky Rozhlas[3], Euranet[4] (European Radio Network) and Radio Regina. Slovensky Rozhlas (Slovak radio) is the Slovakian's national public-service radio broadcaster. Radio Regina is one of the six radio channels of Slovensky Rozhlas. The speech data include daily broadcast news and interviews. In total, we have collected 182.5 hours of speech acoustic data. This speech data are used to develop the STT system, including the acoustic models training and the calculation of the PCA transformation matrix [15]. BN and BC speech data for development and evaluation are collected separately and independently, by the 2012 Quaero STT evaluation organizers.

Text data are collected from the Internet to train language models. The text sources are selected to be able to cover the topics which are often mentioned in the BN and BC. Hence, text data have been collected from 4 sources which are daily newspapers (sources #1 and #4), personal blogs (source #2), weekly magazines of politics, culture and economy (source #3). The information and statistics of the text data downloaded from these sources are summarized in table 1. The text data downloaded from these sources are then normalized in order to keep only clean text. The conversion of the numbers, dates and time into their pronunciations has been also performed.

Table 1. Information and statistics of text data (after normalization) collected for language models training. Sources #1 and #4: daily newspapers. Source #2: personal blogs. Source #3: weekly magazine of politics, culture and economy.

Text sources	# Sentences	# Words
#1. http://zivot.azet.sk	755K	7M
#2. http://blog.sme.sk	940K	11M
#3. http://www.noveslovo.sk	1.0M	15M
#4. http://lesk.cas.sk	3.8M	30M
Total	6.5M	63M

4. N-GRAM LANGUAGE MODELS

N-gram language models (LMs) are trained using the normalized text data. A vocabulary consisting of 439K words has been used in the training of the LMs. The words in the vocabulary are selected as those appearing more than 2 times in the text data for training the LMs. The 1-gram, 2-gram, 3-gram and 4-gram have been trained on the 4 text sources. As mentioned previously, a set of speech data has been collected by the 2012 Quaero STT evaluation organizer. This speech data as well as their manual transcription are available for system development. We make use of the development text, consisting of 20K words, to evaluate our LMs. The out-of-vocabulary (OOV) rate, calculated on this development text, equals 1.64%. The perplexities (PPLs) and the hit rates of the LMs, calculated on the development text, are shown in table 2. Interpolated 2-, 3- and 4-gram LMs are built from the individual n-gram LMs. The interpolation weights, perplexities and hit rates of these LMs are shown in table 2. The perplexities of the interpolated LMs, calculated on the development text, are reduced compared to those of the individual LMs. These interpolated LMs are used in the STT system.

5. PRONUNCIATION DICTIONARY

In a STT system, the pronunciation dictionary makes a link between the language and acoustic models. Indeed, state-of-the-art medium

[3] http://www.rozhlas.sk

[4] http://www.euranet.eu

Table 2. Perplexities, hit rates of the 2-, 3-, 4-gram and the interpolated LMs, calculated on the 20K-word development text. The sum of the interpolation weights of the interpolated LMs, which are mixture LMs of the corresponding n-gram LMs, equals 1.

LM	Weight	PPL	1-gram	2-gram	3-gram	4-gram
Source #4	0.34	1285	26.31	46.24	21.14	6.31
Source #3	0.32	1401	29.81	46.21	18.90	5.08
Source #2	0.21	1551	32.88	45.38	17.28	4.45
Source #1	0.13	1884	38.25	43.60	14.65	3.50
Int 4-gram	-	867	19.07	45.86	25.86	9.20
Int 3-gram	-	883	19.07	45.86	35.07	-
Int 2-gram	-	1059	19.60	80.40	-	-

or large vocabulary STT systems use pronunciation dictionaries as knowledge sources to transcribe individual words into model structures [18]. This is especially the case within a hidden Markov model (HMM) framework where the construction of word HMMs from phoneme models is straightforward. The transcription from words to sequences of phonemes makes use of grapheme-to-phoneme (g-to-p) rules of the language in question. The set of phonemes used in the pronunciation dictionary would be utilized as seed acoustic models (CI models).

To create the pronunciation dictionary, rule-based [19] or data-driven [20] g-to-p approaches can be utilized. Data-driven g-to-p requires sufficient amount of training data to generalize the pronunciations for all the entries of the dictionary. The rule-based g-to-p approach requires the analysis of all phonetic phenomena in Slovak language. In this paper, we use the rule-based g-to-p approach to create the pronunciation dictionary which contains 439K words (entries). The g-to-p rules for Slovak language could be concerning vowels (e.g. *i, ä*), diphthongs (e.g. *ia, ie, iu, ô*), consonants (e.g. *m, f*) or doubled consonants (e.g. *ts, dz*). Otherwise, there are particular rules for voice assimilation, vowel sequence or hard vocal begin and glottal stop [19]. The rule-based g-to-p for Slovak language is, thus, context-dependent and consists of a large number of transcription rules.

We make use of a rule-based g-to-p tool [19], for Slovak language, to transcribe the words in the dictionary into their pronunciations. This tool, called g2p-sk[5], implements all the Slovak g-to-p rules mentioned previously (257 rules). Performance of the phonetic transcription depends on the morphemic and syllabic segmentations [19]. In this tool, these segmentation rules are stored in the built-in dictionaries of the tool. The tool can generate multiple pronunciations for a word, resulting in an average of 1.2 pronunciations/word for the 439K words in the dictionary. The phonetic transcription is context-dependent [19]. After the automatic transcription, phonemes' grouping has been performed, by grouping phonemes with similar acoustic realizations, in order to reduce the number of CI acoustic models. Finally, the dictionary consists of 37 phonemes, including 26 consonants and 11 vowels.

6. STT SYSTEM DEVELOPMENT

6.1. Acoustic feature extraction

6.1.1. Cepstral features

The cepstral feature vector consists of 39 PLP-like (perceptual linear predictive) coefficients [21] derived from a Mel frequency spectrum estimated on the telephone bandwidth (0-8kHz), every 10 ms. Cepstral mean removal and variance normalization are carried out on the basis of speech clusters, obtained after automatic speech segmentation and speaker clustering, resulting in a zero mean and unity variance for each cepstral coefficient. The 39-dimensional acoustic feature vector consists of 12 cepstral coefficients and the log energy, along with the first and second derivative coefficients. A 3-dimensional pitch feature vector (pitch, Δ and $\Delta\Delta$ pitch) is extracted, using the autocorrelation method together with linear interpolation [12], and added to the original PLP features, resulting in a 42-dimensional cepstral feature vector (PLP+F0).

6.1.2. MLP features

The MLP features are generated in two steps. The first step is raw features extraction which constitutes the input layer to a MLP neural network $\mathbb{M}$. In this work, the TRAP-DCT (TempoRAl Pattern - Discrete Cosine Transform) [22] is used as raw features. The TRAP-DCT features are obtained from a 19-band Bark scale spectrogram, using a 30 ms window and a 10 ms offset. A discrete cosine transform (DCT) is applied to 500-ms window of each band from which 25 first DCT coefficients are retained. The retained DCT coefficients are then concatenated together. In total, the raw features have, thus, $19 \times 25 = 475$ DCT coefficients. The raw features are then input to the 4-layer MLP $\mathbb{M}$ [11] with the bottle-neck architecture [22]. The size of the third layer (the bottle-neck) is equal to the desired number of features (39). In a second step, the raw features are processed by the MLP $\mathbb{M}$ and the features are not taken from the output layer of the MLP $\mathbb{M}$ but from the hidden bottle-neck layer and decorrelated by a PCA transformation. The MLP feature vector has finally 39 dimensions. An illustration of MLP (bottle-neck) feature extraction is shown in Fig. 1.

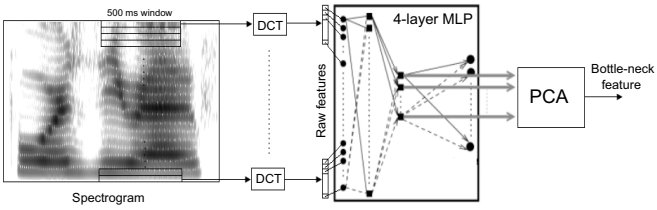


Fig. 1. MLP (bottle-neck) features extraction using a 4-layer MLP neural network. The input features are TRAP-DCT, extracted from 500 ms windows in the subbands of the short-term spectrogram [12, 22]. The bottle-neck features (39-dimensional) are extracted after using a PCA to decorrelate the outputs of the bottle-neck layer.

An interesting aspect of MLP features extraction is that the MLP $\mathbb{M}$ can be trained on the data that is different from the domain of the task [23], while still yielding good generalization performance. This characteristic makes it possible to train the MLP $\mathbb{M}$ on full-resourced languages, for instance English, for using with low-resourced languages. In this respect, the MLP neural network $\mathbb{M}$ is trained on about 645 hours of English broadcast news (BN) data which is readily available at LIMSI. Part of the training data is broadcast data available by the Linguistic Data Consortium (LDC)(Hub4 and TDT corpora [24]). The rest of the data was collected in several projects (mainly TCSTAR, Quaero) and transcribed by partners in them. The audio comes from a variety of news sources (ABC, Skynews, BBC F24 Euronews, ITV1, etc.) and was mostly collected via satellite with some downloaded from the web. Since the amount of data

[5]http://packages.debian.org/en/squeeze/g2p-sk

for training the MLP $\mathbb{M}$ is very large, an efficient training procedure should be implemented. In our work, a simplified training scheme, proposed in [13], was applied for the training. Following this scheme, the training data are randomized and split in three non-overlapping subsets, used in 6 training epochs with fixed learning rates. The first three epochs use only 13% of data, the next two use 26%, the last epoch uses 52% of the data, with the remainder used for cross-validation to monitor the performance. The Quicknet[6] software was used to train the MLP. The MLP has 138 targets, corresponding to the individual states for each phone and one state for the additional pseudo phones (silence, breath, filler-word). The outputs of the MLP were normalized to range between 0 and 1 using the softmax function.

6.1.3. Features' dimension reduction using PCA

The cepstral features (C-dimensional) are augmented with the discriminative features (D-dimensional). The augmented feature vector $\mathbf{y}$ has cumulative dimension (L-dimensional, $L = C + D$) and could contain redundant information for speech recognition. We propose to reduce the dimension of the feature vector $\mathbf{y}$ using principal component analysis (PCA). To reduce the dimension of the augmented feature vector $\mathbf{y}$ by PCA, a transformation matrix $\mathbf{P}$ of $L \times L$ dimensions, whose columns are the principal components, is calculated. The augmented feature vectors $\mathbf{y}$ are then linearly transformed to a lower dimension feature vector $\widehat{\mathbf{y}}$ using a matrix $\widehat{\mathbf{P}}$ of $L \times M$ dimensions ($M < L$), which contains M first principal components, following the equation:

$$\widehat{\mathbf{y}} = \widehat{\mathbf{P}}^T \mathbf{y}$$

where ^{T} denotes the transpose. The M-dimensional feature vector $\widehat{\mathbf{y}}$ is then used in the training and testing of the speech recognition system. To calculate the matrix $\mathbf{P}$, disjoint data, which are not used in train and test, are selected. Augmented feature vectors (L-dimensional) are extracted from these data and are put adjacently in a matrix $\mathbf{Y}$. After that, the matrix $\mathbf{P}$ is calculated from the data matrix $\mathbf{Y}$ by singular value decomposition (SVD) technique [25]. This matrix is calculated once and is the only matrix using for features projection.

6.2. Acoustic models training with MLP+PLP+F0 features

The above mentioned acoustic data (see section 3), consisting of 182.5 hours of acoustic data, is divided into two sets S1 and S2, in respecting of the diffusion time. The first one, S1, consists of 109.5 hours, broadcast in 2011, and the second one, S2, consists of 73 hours of acoustic data, broadcast in 2012. In fact, after partitioning the data into homogenous segments [26], only 92.5 hours of speech from S1 and 60 hours of speech from S2 are retained. In this work, we use speech data from S1 (92.5 hours) for STT system training. The speech data from S2 (60 hours) are used to calculate the PCA projection matrix $\mathbf{P}$ (see section 6.1.3).

Since no transcription of the acoustic data is available, unsupervised training techniques are applied to train context-dependent acoustic models. More specifically, the cross-language transfer technique [8] is applied to create initial CI acoustic models (AMs). The CI AMs are gender-independent (GI). These CI AMs consist of the 37 phonemes presented in the pronunciation dictionary. These initial CI AMs are selected from the CI AMs of the already available STT systems at LIMSI, including Arabic [27], English [28], French [6] and Russian [29] STT systems. The 37 phonemes and their corresponding acoustic sources are shown in table 3. In addition, 3 CI AMs from other languages, which model silence, breath and filler-word, have also been involved, resulting in a total of 40 CI AMs. Indeed, it would be quasi-ideal if the CI AMs could be selectecd from Czech STT system since Czech is the closest language to Slovak. However, we do not have Czech STT system at LIMSI. Furthermore, we have also noticed that the CI AMs are not very important for unsupervised training. Therefore, the CI AMs have been selected from the available STT systems at LIMSI, namely English, French and Russian STT systems.

Table 3. The 37 phonemes in the dictionary and the corresponding source phonemes, taken from existent STT systems (Ar: Arabic, En: English, Fr: French, Ru: Russian). These phonemes are used as the context-independent (CI) acoustic models.

Phonemes	Source phonemes	Phonemes	Source phonemes
b	b (Fr)	l	l (En)
d	d (Fr)	L	ly (Ru)
g	g (Fr)	m	m (En)
p	p (Fr)	n	n (En)
t	t (Fr)	N	ny (Ru)
k	k (Fr)	f	f (Fr)
c	ts (Ru)	v	v (Fr)
C	tS (En)	j	y (En)
$	z (En)	r	r (Ru)
D	dy (Ru)	s	s (Fr)
T	ty (Ru)	z	z (Fr)
h	h (En)	S	S (En)
H	x (Ar)	Z	Z (Fr)
A	^ (En)	o	c (En)
a	a (En)	ó	o (Fr)
E	E (En)	ô	o (En)
é	e (Fr)	u	U (En)
I	I (En)	ú	u (En)
i	i (Fr)		

We apply unsupervised acoustic models training [1] to train context-dependent (CD) AMs, since there are no available transcriptions of the training data. In this respect, these 40 CI and GI AMs are used in the initial step to decode the speech data for training, from S1. It should be noted that the CI AMs, taken from existent STT systems, have been trained with cepstral features (PLP+F0). The hypotheses obtained from the decoding with the CI AMs are used as the ground truth for subsequent decoding iterations. The unsupervised acoustic models training is continued with the following steps:

i. During the second iteration, context-dependent (CD) and GI AMs are trained with the hypotheses obtained in the first iteration.

ii. Gender-dependent (GD) and CD AMs are trained during the third iteration with the hypotheses obtained in the second iteration. GD models are obtained by MAP (*maximum a posteriori*) adapting [30] the context-dependent GI models, using male and female data whose labels are produced after the audio partitioning [26].

iii. The MLP+PLP+F0 features, obtained by concatenating the MLP and PLP+F0 features [10], are used in the fourth iteration. This iteration uses the hypotheses produced from the third iteration with AMs trained with cepstral features (PLP+F0). Context-dependent, GI and GD acoustic models with MLP+PLP+F0 features are obtained during this iteration.

[6] http://www1.icsi.berkeley.edu/Speech/qn.html

iv. The final MLP+PLP+F0 AMs are trained with the hypotheses obtained from the decoding of the training speech data (from S1) using the AMs based on MLP+PLP+F0 features.

In our experiments on the development data, the initial decoding with context-independent (CI) AMs, using PLP+F0 features, gave a WER of 90.8%. With PLP+F0 features, the lowest WER has been obtained with the context-dependent (CD) gender-independent (GI) AMs. This WER equals 43.25%. The WERs obtained with the CD AMs, trained with MLP+PLP+F0 features, are reported in section 7.

6.3. Acoustic models training with reduced features

To reduce the dimension of the MLP+PLP+F0 features with PCA (principal component anaysis), the PCA matrix $\mathbf{P}$ (see section 6.1.3) is calculated and used for features projection. In this respect, disjoint speech data from S2 (60 hours) are utilized to estimate $\mathbf{P}$. After partitioning, there are 3416 speech segments in this set. MLP+PLP+F0 feature vectors are calculated from these speech segments, resulting in a total of a 15.5M feature vector. Following [15], a limited number (13.5K) of MLP+PLP+F0 feature vectors are extracted from 15.5M feature vectors to constitute the data matrix $\mathbf{Y}$ (see section 6.1.3) which is used to calculate the matrix $\mathbf{P}$. The extraction is performed by randomly selecting 4 MLP+PLP+F0 feature vectors from each speech segments. This selection ensures that the matrix $\mathbf{Y}$ is constituted by speech data which cover various environmental and speaking conditions.

It has been shown that 50 is the dimension of the reduced features which gave best performance in speaker verification application [15]. In the current study, the 81-dimensional MLP+PLP+F0 feature vectors are projected into the PCA space $\mathbf{P}$ to create 50-dimensional ($M = 50$) feature vectors. We create also 81-dimensional ($M = 81$) feature vectors in order to assess the effectiveness of the decorrelation and the dimension reduction performed by PCA. Feature vectors created with PCA are denoted as MLP+PLP+F0-PCA features and the specific ones are denoted as MLP+PLP+F0-PCA50 ($M = 50$) and MLP+PLP+F0-PCA81 ($M = 81$) features. Further exhaustive study would be performed to investigate the impact of the reduced features' dimension to the STT system performance. To train the AMs with MLP+PLP+F0-PCA features, the steps mentioned previously (see section 6.2) are repeated. The MLP+PLP+F0-PCA features are used instead of the MLP+PLP+F0 features.

7. EXPERIMENTAL RESULTS

7.1. Word error rate (WER)

The results are reported with the AMs based on MLP+PLP+F0 and MLP+PLP+F0-PCA features since these are the final models of the training process. These AMs gave better performance compared to intermediate AMs. The STT system performance, in term of case-insensitive word error rate (WER) calculated on the development set (3.5 hours of speech), are shown in table 4. These results are obtained with one-pass decoding with several steps. The lattices are rescored using the interpolated 4-gram language model [6]. Silence models of different sizes are tested. The small silence model is 96-Gaussian and the bigger one is 1024-Gaussian. Gender-independent (GI) and gender-dependent (GD) AMs are utilized. It can be observed that:

- Small silence models (96-Gaussian) work better than the bigger ones (1024-Gaussian).
- GI AMs work better than the GD AMs.
- STT system using the MLP+PLP+F0-PCA50 features has a smaller WER compared to the system using the standard MLP+PLP+F0 features, in each comparable condition.
- STT system using the MLP+PLP+F0-PCA81 features has the highest WER, amongst three types of features, in all comparable conditions.

STT experiments are performed with another independent set (4.8 hours) of speech data which was released by the 2012 Quaero STT systems evaluation organizers. The WERs are shown in table 5. The phenomena, observed on the development set, repeat with the 2012 Quaero evaluation set. These results show that the reducing the dimension of the feature vectors from 81 to 50 helps in reducing the WER whereas decorrelating the coefficients of the standard MLP+PLP+F0 feature vectors, without reducing the dimension, does not help to reduce the WER.

7.2. Discussion

In the training of gender-dependent (GD) acoustic models (AMs), labeled male and female data, for adapting the gender-independent (GI) AMs to the GD AMs, have been obtained from a standard audio partitioner [26]. The WERs, obtained with the GD AMs, are not as good as those obtained with the GI AMs. The fact that GD AMs are not effective, compared to the GI AMs, may be due to the fact that there is much more male data than female data in our training data. In this context, the use of MLP+PLP+F0-PCA features helps in reducing the detrimental effect created by the unbalance of GD data.

Further, the fact that smaller silence model works better than larger silence model might be due to the fact that the completely unsupervised training was not converging well for the silence. Further analyses on the effect of silence models to the results should be carried out. On another aspect, using the MLP+PLP+F0-PCA features reduces the recognition time slightly (6% relative in average) while reducing the CMLLR adaptation time by a factor of 3, compared to when using the MLP+PLP+F0 features.

8. CONCLUSION

We have reported the development of a speech-to-text (STT) system for transcribing Slovak broadcast news (BN) and broadcast conversation (BC). Relevant techniques, including cross-language transfer [8] and unsupervised acoustic models training [1], have been utilized during the system development (for training context-dependent acoustic models without any manual transcription of the training data). These techniques, together with the efficient application of N-gram language modeling and rule-based pronunciation modeling [19], have made it possible to develop a state-of-the-art STT system for low-resourced Slovak language. Furthermore, we have proposed to reduce the dimension of the MLP+PLP+F0 features, using principal component analysis (PCA), in order to reduce the redundancy between the MLP and the PLP+F0 features. This features' dimension reduction has made it possible to reduce the WERs as well as the recognition and adaptation time (with CMLLR transform [17]) of the STT system.

Table 4. Word error rates (WERs, in %), calculated on the *development* set (3.5 hours of speech), with MLP+PLP+F0 and MLP+PLP+F0-PCA features. The GI and GD acoustic models (AMs) were tested with two silence models (96- and 1024-Gaussian).

Features / Silence model	MLP+PLP+F0	MLP+PLP+F0-PCA81	MLP+PLP+F0-PCA50
96-Gaussian (GI AMs)	24.53	25.02	**24.33**
1024-Gaussian (GI AMs)	26.10	26.16	**24.90**
96-Gaussian (GD AMs)	25.98	26.00	**24.64**
1024-Gaussian (GD AMs)	28.20	27.70	**25.68**

Table 5. Word error rates (WERs, in %), calculated on the *evaluation* set (4.8 hours of speech), with MLP+PLP+F0 and MLP+PLP+F0-PCA features. The GI and GD acoustic models (AMs) were tested with two silence models (96- and 1024-Gaussian).

Features / Silence model	MLP+PLP+F0	MLP+PLP+F0-PCA81	MLP+PLP+F0-PCA50
96-Gaussian (GI AMs)	29.07	29.79	**29.00**
1024-Gaussian (GI AMs)	31.35	31.45	**29.63**
96-Gaussian (GD AMs)	30.21	30.61	**29.35**
1024-Gaussian (GD AMs)	33.03	32.88	**30.37**

Acknowledgments

This work has been partially financed by OSEO, the French State Agency for Innovation, under the Quaero program. The authors would like to thank Thiago Fraga Da Silva (LIMSI-CNRS), Abdel Messaoudi (Vocapia Research) and Ilya Oparin (LNE, France) for the valuable help during the system development.

9. REFERENCES

[1] L. Lamel, J.-L. Gauvain, and G. Adda, "Unsupervised acoustic model training," in *Proc. IEEE ICASSP'02*, Orlando, USA, May 2002, pp. 877–880.

[2] T. Fraga-Silva, V.-B. Le, L. Lamel, and J.-L. Gauvain, "Incorporating MLP features in the unsupervised training process," in *3rd Workshop on Spoken Language Technologies for Under-resourced Languages*, Cape Town, South Africa, May 2012, pp. 24–28.

[3] I. Oparin, L. Lamel, and J.-L. Gauvain, "Rapid development of a Latvian speech-to-text system," in *Proc. IEEE ICASSP'13*, Vancouver, Canada, May 2013, pp. 7309–7313.

[4] A. Roy, L. Lamel, T. Fraga da Silva, J.-L. Gauvain, and I. Oparin, "Some issues affecting the transcription of Hungarian broadcast audio," in *Proc. INTERSPEECH'13*, Lyon, France, August 2013, pp. 3102–3106.

[5] P.C. Woodland, T. Hain, S.E. Johnson, T.R. Niesler, A. Tuerk, and S.J. Young, "Experiments in broadcast news transcription," in *Proc. IEEE ICASSP'98*, Seattle, WA, USA, May 1998, pp. 909–912.

[6] J.-L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, and H. Schwenk, "Where are we in transcribing French broadcast news?," in *Proc. INTERSPEECH'05*, Lisbon, Portugal, September 2005, pp. 1665–1668.

[7] J. Nouza, J. Silovsky, J. Zdansky, P. Cerva, M. Kroul, and J. Chaloupka, "Czech-to-Slovak adapted broadcast news transcription system," in *Proc. INTERSPEECH'08*, Brisbane, Australia, Sep. 2008, pp. 2683–2686.

[8] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, pp. 31–51, August 2001.

[9] H. Bourlard et al., "Current trends in multilingual speech processing," *Sadhana*, vol. 35, pp. 885–915, October 2011.

[10] N. Morgan et al., "Pushing the envelope - Aside," *IEEE Signal Processing Magazine*, vol. 22, pp. 81–88, September 2005.

[11] P. Fousek, L. Lamel, and J.-L. Gauvain, "Transcribing broadcast data using MLP features," in *Proc. INTERSPEECH'08*, Brisbane, Australia, September 2008, pp. 1433–1436.

[12] L. Lamel, J.-L. Gauvain, V.-B. Le, I. Oparin, and S. Meng, "Improved models for Mandarin speech-to-text transcription," in *Proc. IEEE ICASSP'11*, Prague, Czech Republic, May 2011, pp. 4660–4663.

[13] Q. Zhu, A. Stolcke, B.-Y. Chen, and N. Morgan, "Using MLP features in SRI's conversational speech recognition system," in *Proc. INTERSPEECH'05*, Lisbon, Portugal, September 2005, pp. 2141–2144.

[14] T. Schultz, N.-T. Vu, and T. Schlippe, "GlobalPhone: a multilingual text & speech database in 20 languages," in *Proc. IEEE ICASSP'13*, Vancouver, Canada, May 2013, pp. 8126–8130.

[15] C.-T. Do, C. Barras, V.-B. Le, and A.-K. Sarkar, "Augmenting short-term cepstral features with long-term discriminative features for speaker verification of telephone data," in *Proc. INTERSPEECH'13*, Lyon, France, August 2013, pp. 2484–2488.

[16] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–186, April 1995.

[17] V.-V. Digalakis, D. Rtischev, and L.G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 357–366, September 1995.

[18] T. Hain, "Implicit modelling of pronunciation variation in automatic speech recognition," *Speech Comm.*, vol. 46, pp. 171–188, June 2005.

[19] J. Ivanecky, "Analysis of the rule based phonetic transcription technique applied to the Slovak language," in *Proc. SLOVKO'05 - 3rd Int. Sem. on Computer Treatment of Slavic and East European Languages*, Bratislava, Slovakia, November 2005, pp. 130–136.

[20] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, pp. 434–451, May 2008.

[21] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, pp. 1738–1752, April 1990.

[22] F. Grezl and P. Fousek, "Optimizing bottle-neck features for LVCSR," in *Proc. IEEE ICASSP'08*, Las Vegas, USA, March-April 2008, pp. 4729–4732.

[23] S. Sivadas and H. Hermansky, "On use of task independent training data in Tandem feature extraction," in *Proc. IEEE ICASSP'04*, Montreal, Canada, May 2004, pp. 541–544.

[24] C. Cieri, D. Graff, M. Liberman, N. Martey, and S. Strassel, "The TDT-2 text and speech corpus," in *Proc. 1999 DARPA Broadcast News Workshop*, Virginia, USA, February 1999.

[25] I. T. Jolliffe, *Principal component analysis*, Springer-Verlag, 2nd Eds, 2002.

[26] J.-L. Gauvain, L. Lamel, and G. Adda, "Partitioning and transcription of broadcast news data," in *Proc. ISCA ICSLP'98*, Sydney, Australia, December 1998, pp. 1335–1338.

[27] L. Lamel, A. Messaoudi, and J.-L. Gauvain, "Automatic speech-to-text transcription in Arabic," *ACM Transactions on Asian Language Processing*, vol. 8, no. 4, December 2009.

[28] R. Prasad et al., "The 2004 BBN/LIMSI 20xRT English conversational telephone speech recognition system," in *Proc. INTERSPEECH'05*, Lisbon, Portugal, September 2005, pp. 1645–1648.

[29] L. Lamel et al., "Transcription of Russian conversational speech," in *Proc. SLTU'12 3rd Intl. Work. on Spoken Lang. Tech. for Under-resourced Lang.*, Cape Town, South Africa, May 2012, pp. 162–167.

[30] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech & Audio Process.*, vol. 2, pp. 291–298, April 1994.