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

International Workshop on
Spoken Language Translation 2011
San Francisco, CA, USA
December 8-9, 2011

Speech Recognition for Machine Translation in Quaero

Lori Lamel[1], *Sandrine Courcinous*[4], *Julien Despres*[4], *Jean-Luc Gauvain*[1], *Yvan Josse*[4], *Kevin Kilgour*[2], *Florian Kraft*[2], *Viet Bac Le*[1,4], *Hermann Ney*[3], *Markus Nußbaum-Thom*[3], *Ilya Oparin*[1], *Tim Schlippe*[2], *Ralf Schlüter*[3], *Tanja Schultz*[2], *Thiago Fraga da Silva*[1], *Sebastian Stüker*[2], *Martin Sundermeyer*[3], *Bianca Vieru*[4], *Ngoc Thang Vu*[2], *Alexander Waibel*[2], *Cécile Woehrling*[4]

[1] CNRS-LIMSI, Orsay, France (www.limsi.fr/tlp)
[2] Karlsruhe Institute of Technology, Karlsruhe, Germany (www.informatik.kit.edu/interact)
[3] RWTH Aachen University, Aachen, Germany (www-i6.informatik.rwth-aachen.de)
[4]Vocapia Research, Orsay, France (www.vocapia.com)

Abstract

This paper describes the speech-to-text systems used to provide automatic transcriptions used in the Quaero 2010 evaluation of Machine Translation from speech. Quaero (www.quaero.org) is a large research and industrial innovation program focusing on technologies for automatic analysis and classification of multimedia and multilingual documents. The ASR transcript is the result of a Rover combination of systems from three teams (KIT, RWTH, LIMSI+VR) for the French and German languages. The case-sensitive word error rates (WER) of the combined systems were respectively 20.8% and 18.1% on the 2010 evaluation data, relative WER reductions of 14.6% and 17.4% respectively over the best component system.

1. Introduction

This paper describes the speech recognition systems used for the Quaero 2010 evaluation of Machine Translation from speech. An accompanying paper [1] describes the Machine Translation systems evaluated, comparing MT performance on the automatic and reference transcriptions.

The Quaero[1] is a large research and industrial innovation program focusing on technologies for automatic analysis and classification of multimedia and multilingual documents. The program has two projects devoted to research and common resources led by academic partners, and five application projects led by industrial partners. The core technologies are developed within the Quaero Core Technology Cluster (CTC) project which has the main research goals improving the state-of-the-art in automatic multimedia document structuring for indexing, by developing and evaluating the underlying techniques and models. The core technologies are for text processing, translation, audio and speech processing, image and video processing, data protection, cross-modal processing, and search and navigation methods for multimedia and multilingual documents. Evaluation campaigns have been held annually since the start of Quaero covering more than 30 technologies, including speech-to-text (STT) and spoken language translation (SLT).

Four partners contribute to the STT task in Quaero: CNRS-LIMSI, KIT, RWTH and Vocapia Research (VR). At the program start focused on the 3 primary Quaero languages: French, German and English; with two additional languages added each year. In 2010, STT was evaluated for 7 languages (adding Spanish, Russian, Greek, Polish) with Italian and Portuguese introduced in 2011. For STT, the evaluations are organized by the LNE (Laboratoire National de M'etrologie et d'Essais[2]) coordinated by the DGA (Délégation Générale pour l'armement[3]) who also organizes the MT evaluations.

The 2011 Quaero spoken language translation internal evaluation campaign [1] addressed bi-directional translation between the French-German languages. SLT performance was compared using both manual references and automatic transcripts of the spoken text as input data. The automatic transcripts were produced by a Rover combination of the best submission from each site to the Quaero 2010 ASR evaluation as described in Section 6. The SLT evaluation campaign was organized and run by the DGA, so as to compare the different approaches taken by the four SLT participating partners RWTH, KIT, LIMSI and SYSTRAN.

2. Quaero Speech-to-text Systems

In the following sections the individual STT systems from the 4 participating sites are described. The systems employ most techniques found in today's state-of-the-art systems, and research addresses many leading edge topics. Comparable results are obtained across sites for mature systems.

Table 1 summarizes the characteristics of Quaero 2010 evaluation data for French and German. For 2010, the STT development data was the 2009 evaluation data, containing 3.5 hours of data for each language, primarily Webdata, complemented with Broadcast News (BN) and 30 minutes of European Parliament Plenary Sessions (EPPS) data. The 2010 evaluation data contains roughly 30% Broadcast News and 70% varied broadcast data referred to as Broadcast Conversation (BC).

The lexical coverage of the recognition vocabulary is an important factor in a STT system, since any unknown (referred to as out-of-vocabulary or OOV) words will result in a recognition error. The vocabulary sizes range from 65k to 300k words, with some sites using multiple word lists. The word lists are selected either using frequency cut-offs or unigram interpolation. In general the OOV rates range from about 0.5 to 2%. The systems represent the pronunciations with sets of 35 to 50 phone symbols, and generate the pronunciations with different methods. Some systems use rule-based grapheme to phoneme conversion, others statistical methods, or a combination of the two, often with a list of (possibly manually verified) exceptions. Most phone sets include pseudo phones for silence and non-speech sounds. There are typically 1.1 to 1.3 pronunciations per word.

[1] www.quaero.org

[2] www.lne.fr

[3] www.defense.gouv.fr/dga

Table 1: *Summary of Quaero 2010 evaluation data.*

Language	Broadcast News	Broadcast Conversation	Total
French	0:53	2:07	3:00
German	1:20	2:14	3:34

The acoustic models are trained on several hundreds of hours of audio data from a variety of sources, mainly from the Linguistic Data Consortium (LDC) or from previous European projects. In addition to data transcribed in Quaero, French data from BREF, ESTER and EPPS (European Parliament Plenary Sessions) and the German data from GlobalPhon, Verbmobil, LBW, WDR, Mainz, Zeit were used. Not all sites have access to all data, and the sites relied on untranscribed data for some languages.

The language models are trained on over a billion words of texts, comprised of assorted newspaper and newswire texts, including the LDC Gigaword corpora, as well as Webtexts, EPPS documents, commercial transcript s and closed captions. Audio transcripts represent only a small percentage of the training material, with only between several hundred thousand to several million words. The language models are quite large, typically containing over 400M 4-grams. The perplexity of the 2010 development data is about 130 for French and 250 for German.

3. KIT STT systems

3.1. KIT 2010 German STT system

All speech recognition experiments described in the following were performed with the help of the Janus Recognition Toolkit (JRTk) and the Ibis single pass decoder [3].

3.1.1. Front-End and Acoustic Model Training

Two different front-ends were applied: The warped minimum variance distortionless response (WMVDR) approach and the conventional (Mel-frequency Cepstral Coefficients) MFCC approach. The front-end uses a 42-dimensional feature space with linear discriminant analysis and a global semi-tied covariance (STC) transform [4] with utterance-based cepstral mean and variance normalization. The 42-dimensional feature space is based on 20 cepstral coefficients for the MVDR system and on 13 cepstral coefficients for the MFCC system.

The training setup was based on last years evaluation system. The following training material was used: Quaero 2009 training data (6 hours EPPS, 14 hours web data) and development data (13 hours), Quaero 2010 training data set (51 hours), Verbmobil (67 hours), recordings of the Landtag Baden-Wuerttemberg (123 hours), Tagesschau (17 hours), isl-database (16 hours), Globalphone (19 hours), in-house lecture and talk recordings (26 hours),

All the acoustic data is in 16 kHz, 16 bit quality. Acoustic model training was performed with fixed state alignments and Vocal Tract Length Normalization (VTLN) factors, which were obtained using the 2009 evaluation system. The system uses left-to-right hidden Markov Models (HMM)s without state skipping with three HMM states per phoneme. In additional to the 2009 setup with 2000 distributions and codebooks with up to 128 Gaussians per model using the MVDR front-end, the same setup with the MFCC front-end and also new systems with 4000 distributions for both front-ends were trained. The adapted gender-independent acoustic model training (given the vocal tract normalization values for each speaker by the previous system) can be outlined by the following training sequence: training of the linear discriminant analysis matrix, extraction of samples, incremental growing of Gaussians, training of one global STC matrix, second extraction of samples, second incremental growing of Gaussians, three iterations of Viterbi training and three iterations of FSA-SAT speaker adaptive training. For the 4000 distribution systems the second incremental growing of Gaussians was skipped, since gains couldn't be seen from doing so in other systems.

Table 2: *German text sources used by KIT*

Corpus	*Wordcount*
Transcripts of the Quaero 2009 training data	$130k$
Plenary protocols of the 12th and 13th Baden-Wurttemberg state parliaments.	$12\,538k$
German broadcast news	$863k$ $108\,250k$
Webdumps 2006-2007	$358\,138k$ $16\,319k$ $343\,032k$
Verbmobil text	$641k$
EPPS transcripts (≤ 2006)	$24\,071k$
German ODP webcrawl; only sites archived before 2008 on archive.org	$116\,852k$
German conversational text	$78k$
Quaero 2010 training transcripts	$466k$
Quaero training texts	$747\,573k$
Google N-grams texts	-
total	$1\,612\,099k$

Table 3: *KIT WERs on Quaero German 2010 dev set*

ID	pass	AM	LM	WER in % (ci/cs)
S	Segm.	2k MVDR	LM1	35.5 / 36.4
A	1st	4k MVDR	LM1	30.0 / 31.2
B	1st	4k MFCC	LM1	29.8 / 30.9
C	1st	2k MVDR	LM1	30.8 / 31.9
D	1st	2k MFCC	LM1	31.2 / 32.3
E	CNC A-D			28.3 / 29.4
F	2nd	4k MVDR	LM2	26.8 / 28.0
G	2nd	4k MFCC	LM2	27.0 / 28.0
H	2nd	2k MVDR	LM2	27.7 / 28.8
I	2nd	2k MFCC	LM2	27.9 / 29.0
J	CNC F-I			26.1 / 27.2

3.1.2. Language Model

A language model for the German evaluation system was built from the text sources listed in Table 2. The resulting 10GByte LM contained 31.7M 2-grams, 91.9M 3-grams, 160.4M 4-grams and was reduced to 2.3GByte when stored in an easy to load memory mapped format.

3.1.3. Experiments, Results and Decoding Strategy

This section presents experiments and results on the Quaero development and evaluation data sets. After a segmentation pass and speaker clustering for each of the MVDR and MFCC front-ends, both setups with 2000 and 4000 distributions were decoded using the speaker-independent acoustic models. The result of a confusion network system combination (CNC) [5] applied on all four

Table 4: *KIT WERs on Quaero German 2010 evaluation set*

ID	pass	AM	LM	WER in % (ci/cs)
S	Segm.	2k MVDR	LM1	33.2 / 34.1
A	1st	4k MVDR	LM1	28.1 / 29.3
B	1st	4k MFCC	LM1	28.3 / 29.6
C	1st	2k MVDR	LM1	29.2 / 30.4
D	1st	2k MFCC	LM1	29.8 / 31.0
E	CNC A-D			26.8 / 28.0
F	2nd	4k MVDR	LM2	25.3 / 26.5
G	2nd	4k MFCC	LM2	25.7 / 26.8
H	2nd	2k MVDR	LM2	26.3 / 27.5
I	2nd	2k MFCC	LM2	26.5 / 27.7
J	CNC F-I			24.6 / 25.7
K	comp.			24.1 / 25.2

systems was used to adapt the 2nd pass systems with incremental VTLN adaptation, constrained MLLR and MLLR. In the second pass speaker adapted FSA-SAT models and a bigger language model were used. Finally the four 2nd pass systems were combined again using CNC combination and compound merging was applied on top.

3.2. KIT 2010 French STT system

This section describes the KIT STT system in the Quaero 2010 evaluation for the French language.

3.2.1. Segmentation and Clustering

The segmentation part is implemented in a 2-step approach. In the first step, an HMM-based segmenter is applied, which discriminates speech from events, namely noises, silences and music. To improve the performance of the segmenter on conversational shows which contain of a large amount of speaker changes, back-channel and quick turn-taking behavior, the HMM-segmentation is followed by a speaker change detection post-processing step. Since such turn-changes are more likely to cause longer speech segments, the speaker change detection was only done for segments longer than 5 seconds. The clustering part generates a Gaussian Mixture Model (GMM) for each speech segment by applying MAP adaptation on a tied GMM. Then a hierarchical bottom-up clustering is performed. The closeness relation between two clusters is defined by the Generalized Likelihood Ratio. After merging the two closest clusters, the distances between the merged and all other clusters are updated. The clustering procedure continues until a stopping criteria is met based on Bayesian Information Criterion.

3.2.2. Acoustic Model Training

Approximately 330 hours of audio data were used to train the acoustic model. They are from WEBDATA (26 hrs), ESTER 1 and 2 (164 hrs), EPPS (4 hrs), BREF [2] (80 hrs)and Quaero 2010 training data (60 hrs). Two different kinds of phoneme sets PS1 and PS2 were used for training. PS1 is a modified version of the French GlobalPhone [6] phoneme set which consists of 35 phonemes and PS2 is a version of a phoneme set provided by Vocapia that consists only of 32 phonemes. Furthermore, two different kinds of acoustic front-ends were used. One is based on the traditional Mel-frequency Cepstral Coefficients (MFCC) and the other one on the warped minimum variance distortionless response (WMVDR). Both front-ends provide features every 10ms. For the WMVDR front-end, a model order of 30 and the first 20 cepstral coefficients was used, while the

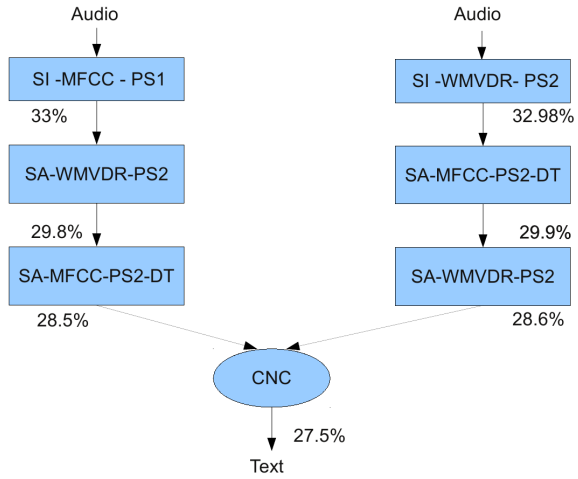

Figure 1: *The decoding strategy for the KIT French STT system.*

MFCC front-end used 13 cepstral coefficients. The mean and variance of the cepstral coefficients were normalized on a per-utterance basis. For both front-ends, seven adjacent frames were combined into one single feature vector. The resulting feature vectors were then reduced to 42 dimensions using linear discriminant analysis (LDA). All models are fully-continuous quinphone systems that use 12,000 codebooks. They were trained using incremental splitting of Gaussians training, followed by 3 iterations of Viterbi training. For all models, we used one global semi-tied covariance (STC) matrix after LDA as well as Vocal Tract Length Normalization (VTLN). In addition, the MLLR speaker adaptive training (SAT) was applied on top. Furthermore, in order to improve the acoustic models, a boosted Maximum Mutual Information Estimation (bMMIE) training was applied after SAT training. This all resulted in five acoustic models: PS1-MFCC, PS2-MFCC-VTLN-SAT, PS2-WMVDR, PS2-WMVDR-VTLN-SAT and PS2-MFCC-VTLN-SAT+bMMIE.

3.2.3. Language Model and Pronunciation Dictionary

In addition to the text data from the Quaero training transcriptions, texts from ESTER 1 and 2, BREF, Hansard, GlobalPhone, French Gigaword, Europarl as well as crawled archives from online news, forums and blogs were used. Additionally, text data were collected with the help of crawling and filter scripts provided by our *Rapid Language Adaptation Toolkit* [7]. With the Quaero training transcriptions and the additional text corpora, LMs which contain the whole vocabulary of the transcriptions were built. Supplemental vocabulary was selected from the additional text material by selecting frequent words which are not in the transcriptions. Based on LM strategies shown in [8], individual 3- and 4-gram LMs were trained with all texts using the SRI Language Modeling Toolkit [9]. The LMs that were used in the systems were created by interpolating the individual models. The interpolation weights were tuned on the 2010 Quaero development set by minimizing the perplexity (PPL) of the model. A 4-gram LM with totally 170k words and a PPL of 181.5 worked out to result in lowest word error rate.

For the training procedure, a dictionary which contains hesitations, fragments, non-human noise and human noise in addition to pronunciation variations of the words was used. The dictionary which was employed for decoding included only non-human noise and pronunciation variants. GlobalPhone and Lexique 3 dictionaries provided different French pronunciations. Missing pronunciations were generated with Sequitur G2P [10].

3.2.4. Decoding strategy and Results

Figure 1 illustrates all details of the decoding strategy and its result on the 2010 Quaero development set. The acoustic models of the actual pass are adapted on the output from the previous pass using Maximum Likelihood Linear Regression (MLLR), Vocal Tract Length Normalization (VTLN) and Feature Space Adaptation (FSA). The CNC which combines the word lattices of the previous passes was also applied. The final system has 27.5% WER on the 2010 Quaero development set.

4. LIMSI/Vocapia STT Systems

LIMSI and Vocapia research jointly developed speech-to-text systems for the Quaero 2010 evaluation. The transcription systems make use of statistical modeling techniques described in [11], which gives details for the English broadcast news system. The acoustic and language models and pronunciation dictionaries are language dependent [12], and trained on large audio and text corpora. Speech decoding is carried out in one or more passes with a statistical n-gram language models. In this section the specific characteristics of the French and German systems used for system combination for the 2011 Quaero Machine Translation from speech evaluation are described.

The first step in processing an audio document is to segment and partition the data, identify the portions containing speech data to be transcribed [13] and associating segment cluster labels, where each segment cluster ideally represents one speaker. Word decoding is carried out in one decoding pass for German and two decoding passes for French. Each decoding pass produces a word lattice with cross-word, word-position dependent acoustic models, followed by consensus decoding with a 4-gram language model and pronunciation probabilities. For French, unsupervised acoustic model adaptation is performed for each segment cluster using the CMLLR and MLLR [14], and the lattices produced are rescored by the neural network LM interpolated with a 4-gram back-off LM.

4.1. Acoustic features

Two types of acoustic features are used. The first set are PLP-like [15] and consist of 39 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band every 10ms. This analysis has been used in all LIMSI STT systems since 1996 [11]. Cepstral mean removal and variance normalization are carried out on a segment-cluster basis, resulting in a zero mean and unity variance for each cepstral coefficient. A 3-dimensional pitch feature vector (pitch, Δ and $\Delta\Delta$ pitch) is added to the original PLP one, resulting in a 42-dimension feature vector (PLP+f0).

The second type are probabilistic features produced by Multi Layer Perceptron (MLP), which have been shown improve system performance when concatenated with cepstral features [16]. Experiments with alternate MLP features have shown that the TRAP-DCT (TD) features have comparable performance to the warped linear predictive temporal patterns (wLP) but are much cheaper to obtain[17]. These 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 each band (the first 25 DCT coefficients are kept) resulting in 475 raw features. features, which are the input to a 4-layer MLP with the bottleneck architecture [18]. The size of the third layer (the bottleneck) is equal to the desired number of features (39). In the second step, the *raw features* are processed by the MLP and the features are not taken from the output layer of the MLP but from the "bottle-neck" hidden layer and decorrelated by a PCA transformation. The second feature vector has 81 parameters resulting from the concatenation of the MLP and PLP and pitch features (MLP+PLP+f0).

Table 5: *Summary of the French language model training texts for the Quaero 2010 evaluation (LIMSI+VR, after normalization).*

French sources	*Epoch*	*#Words*
Press agencies	1994-2005	579M
Press agencies	2006-2007	118M
Google news	2008-2009	173M
Newspapers	1989-2005	609M
Newspapers	2006-2008	86M
Audio transcripts1	2002-2008	1.5M
Audio transcripts 2	1994-2008	7M
Fast transcriptions	1997-2001	94M
Canadian newspaper (La Presse)	2000-2007	206M
Various web data	2000-2004	77M
Various web data	2006-2008	94M
Quaero transcriptions	2008-2010	1.2M
QRTXT	2008-2010	44M
Google News	2010	84M
Internal data newswire	2010	33M

An MLP network was trained for French using the simplified training scheme proposed in [19] on about 300 hours of data from a variety of broadcast sources. The training data are randomized and split in three non-overlapping subsets, used in 6 training epochs with fixed learning rates. The first 3 epochs use only 13% of data, the next 2 use 26%, the last epoch uses 52% of the data, with the remainder used for cross-validation to monitor performance. The MLP has 105 targets, corresponding to the individual states for each phone and one state for the additional pseudo phones (silence, breath, filler).

4.2. Acoustic models

As in [11] the acoustic models are tied-state, left-to-right 3-state HMMs with Gaussian mixture observation densities (typically 32 components). The triphone-based phone models are word-independent, but position-dependent. The states are tied by means of a decision tree to reduce model size and increase triphone coverage. The acoustic models are gender-dependent and speaker-adaptive trained (SAT). Silence is modeled by a single state with 1024 Gaussians.

The acoustic models are trained on the data distributed in Quaero (50-100 hours per language) as well as on data from other sources mainly from previous European or national projects. There were about 300 hours of audio training data for French, and cover 17k phone contexts. The French acoustic models and are discriminatively MMI trained and use probabilistic features based on bottleneck multi-layer perceptrons (MLP) and modified TRAP-DCT features. Combined with classical PLP features, these probabilistic features significantly reduce the word error rate. The German acoustic models use PLP+f0 features and are ML trained on about 80 hours of audio data, and cover 19k phone contexts.

4.3. Language models

Concerning language modeling, the standard n-gram LMs used for both decoding and lattice rescoring are obtained by interpolating unpruned component LMs trained on subsets of the text corpus. Language model training is performed with LIMSI STK toolkit which allows efficient handling of huge language models without any pruning or cutoff. Thus all information in the training data is

Table 6: *Data selection for French NNLM training (LIMSI+VR).*

Corpus	*#sent*	*#word #*	*Weight*	*Sampl./cnt*
BC trans 2008-2010	42k	1.2M	0.26	1.0 / 1.2M
BN qtr 1997-2001	373k	104M	0.19	0.07 / 7.3M
BN trans 1994-2008	315k	8.1M	0.12	0.7 / 5.7M
newspaper 2006-2010	580k	146M	0.07	0.01 / 1.5M

Table 7: *WER of the LIMSI+VR French 4-gram LMs and NNLMs on the Quaero 2010 STT development data.*

LM	*Perplexity*	*WER*
4-gram	94	19.20%
4-gram+NNLMs	92	18.83%

kept, even though there are over 2 billion words. Additional Neural network LMs (NNLMs) used for final lattice rescoring are developed to make use of continuous representation of words, instead of the discrete space in conventional N-gram LMs.

For French two recognition vocabularies were used, one containing 65k words and the other 200k words. The out-of-vocabulary (OOV) rates of the dev data are 1.1% and 0.5% respectively. In total over 2 billion words of texts were used for language modeling. The data come from different sources (Web, newswire, newspaper, detailed and fast manual transcripts), as shown in Table 5. 2-gram, 3-gram and 4-gram language models were developed using Kneser-Ney discounting. The pronunciation lexicon make use of a 35-phone set (3 of which are used for silence, filler words, and breath noises). Baseform pronunciations for the missing words are generated using a grapheme-to-phoneme conversion tool, and alternative pronunciations are added semi-automatically. The 200k word lexicon has 268k pronunciations. The most frequent inflected forms were verified to ensure systematic pronunciations. Pronunciation probabilities are estimated from the observed frequencies in the training data resulting from forced alignment, with a smoothing to account for unobserved pronunciations.

4.4. Neural Network Language models

In contrast to conventional N-gram LMs in which words are represented in a discrete space, Neural network LMs (NNLMs) make use of continuous-space representation of words, which enables a better estimation of unseen N-grams. The neural network deals with two tasks: projection of words with history to continuous space and calculation of LM probabilities for the given history. NNLMs have been shown to improve over the N-gram baseline for different languages and tasks [21].

For the French system four different neural networks were generated with different number of nodes in the hidden layer. The networks vary in the size of the hidden layer (500, 450, 500, 430), and the projection size of P-dimensional continuous space (300, 250, 200, 220). Three previous words form an input to the NN, and the 12k most frequent words are used as a shortlist to estimate the probabilities at the output layer as described in [20, 21].

Since it is not feasible to train a NNLM on all the available texts, the data used to train the NNLMs was selected according to the interpolation weights of the component N-gram LMs in the baseline N-gram LM. This data were downsampled in order to train NNLMs in reasonable time. Only the top four corpora according to N-gram LM interpolation weights were used to train the French NNLMs. The sampling parameters used are given in Table 6. The individual NNLMs were subsequently interpolated together with the baseline 4-gram LMs in order to form the final LM. These LMs are used to rescore lattices generated with conventional N-gram LMs.

Performance results for French with the 4-gram LM and the NNLM on Quaero 2010 STT development data are shown in Table 7. Improvements in both perplexity and word error rate (WER) are observed and even though the perplexity reduction is not particularly impressive, there is a WER reduction of almost 0.4%.

Table 8: *LIMSI+VR German language model training corpus: epoch, total (distinct) number word counts.*

Source	*Epoch*	*#Words (#types)*
die Zeit	01/1946-05/2009	281M (1.4M)
Quaero forum die Zeit	10/2007-12/2009	16M (316K)
Quaero blogs	02/2007-12/2009	21M (364K)
Quaero web	05/2009-02/2010	43M (413K)
der Spiegel	01/1947-05/2009	209M (776K)
google news Switzerland	12/2007-02/2010	159M (511K)
google news Austria	12/2007-02/2010	192M (621K)
google news Germany	12/2007-02/2010	393M (753K)
Quaero transcriptions	08/2008-02/2010	541K (38K)
Other transcriptions	01/1999-01/2002	240K (30K)

Table 9: *Case-insensitive WER of the LIMSI+VR French and German STT systems on the Quaero 2010 dev and eval data.*

WER (%)	French	German
2010 dev	19.5	21.0
2010 eval	19.0	19.9

4.5. German language models

The German language models were trained on a 1.3 billion word text corpus, with a total of 1.7M distinct forms after normalization. The text sources are shown in Table 8 As for all languages, individual N-gram language models are built on each source using the Kneser-Ney smoothing algorithm and then interpolated. The interpolation coefficient were automatically chosen using the EM algorithm. The German recognition lexicon contains 300K words selected by interpolation of unigram models so as to minimize the perplexity on the Quaero development data. The OOV rate is 0.6% on the dev data. The pronunciations are represented with 49 phones, and were derived from the Celex master dictionary and completed with pronunciations from Vocapia's statistical G2P system. The pronunciation probabilities determined from alignment of the acoustic training data.

4.6. Results

Table 9 reports the case-insensitive WER of the LIMSI+VR French and German STT systems on the Quaero 2010 dev and eval data.

5. RWTH STT Systems

In the following, a short description of the French and German RWTH ASR systems is given, which were submitted for the Quaero evaluation 2010.

Table 10: *List of corpora sets used by RWTH for acoustic training*

	corpus	duration	type
DE	Quaero	50 h	BC
FR	Quaero	86 h	BC
	Ester1	43 h	BN
	Ester2	99 h	BN

5.1. Baseline acoustic modeling

For all languages, several subsystems were trained, based on the Mel-Frequency Cepstral Coefficients (MFCC).

Augmenting the MFCC features by a voicedness feature and applying a sliding window of size 9, 154-dimensional feature vectors were obtained that were projected down to 45 components using an LDA transformation. To introduce variability between the resulting subsystems and thereby improve the final system combination step, Perceptual Linear Predictive (PLP) features were also extracted in an analogous manner.

Then phone-posterior-based features, estimated using a multilayer perceptron (MLP), were appended. Regarding the topology, competing approaches included a hierarchical processing (H-MLP) as well as the introduction of a hidden 'bottleneck' layer. This layer corresponds to a dimensionality reduction, as not the outputs of the network are used as tandem features directly but the outputs of this lower-dimensional hidden layer. Both ideas were successfully combined in [22], resulting in the hierarchical bottleneck features (HBN-MLP). For our systems, both H-MLP and HBN-MLP features were used.

The neural networks were trained by feeding the Multi-resolutional RASTA (MRASTA) features as inputs and the phone-posterior probabilities, computed by a forced alignment of the acoustic training data, as desired outputs. As a last step, the MLP features were decorrelated by a PCA transformation. This also allows an additional dimensionality reduction in case of the H-MLPs.

For all systems, acoustic modeling was based on across-word triphone states represented by left-to-right three state Hidden Markov Models (HMMs). Reducing the number of triphone states to 4,500 generalized states via decision tree clustering, the emission probabilities of the remaining states were modeled by Gaussian mixtures with a globally shared covariance matrix and a total of 1 million densities. Baseline acoustic training was performed using Maximum Likelihood with the Viterbi approximation.

Table 10 gives a list of different sets of acoustic training corpora for the individual languages.

5.2. Speaker normalization and adaptation

For most of our subsystems, MFCC and PLP features were normalized using Vocal Tract Length Normalization (VTLN).

To compensate for speaker variation, the Constrained Maximum Likelihood Linear Regression (CMLLR) technique was used in training and recognition. The adaptation matrices were estimated based on alignments computed using single Gaussians (*simple target model* as in [23]) which in general gives better results than full mixture models. For CMLLR, a two-pass recognition setup is necessary.

5.3. Discriminative training

To sharpen acoustic models, discriminative training was applied. Lattices were computed using the current best acoustic models for each language. Based on these lattices, the Minimum Phone Error (MPE) training criterion was optimized[24].

Table 11: *RWTH language model training data for each language.*

	Corpus	*# running words*	*Type*
FR	Gigaword	837 M	newspaper
	Quaero	262 M	blog+news
	Web data	248 M	archives+news
	transcriptions	3 M	BC+BN
DE	Quaero	228 M	blog+news
	Web data	306 M	archives+news
	transcriptions	0.5 M	BC+BN

5.4. Language modeling

Based on the available training data, 4-gram language models (LMs) were estimated for each language using [9], smoothed by the Modified Kneser-Ney method. The LM data was partitioned into blocks, estimating n-gram probabilities for each block individually. The resulting LMs were linearly interpolated while optimizing the perplexity on a holdout data set.

Table 11 gives an overview of the training material that was used. For the 2010 evaluation, substantial amounts of in-domain text data downloaded from web blogs were distributed to all participants. Because of the comparatively small-sized Gigaword corpus for French the LM data were extended by text resources from the web by crawling RSS news feeds and web archives.

Since German is a high inflective language the German language model was enriched by sublexical fragments to reduce the OOV rate and minimize perplexity. Words were decomposed using a statistical data-driven tool. Then a standard n-gram LM was estimated on the decomposed text. Perplexity values were observed of 269, and 131 for German, and French respectively. For a 200 k vocabulary the OOV rate measured for French is around 0.5 % while German has still a higher OOV rate of 1.13%.

5.5. Pronunciation modeling

For all languages, grapheme-to-phoneme models were trained for the creation of pronunciations not found in the baseline lexica, as described in [10].

5.6. Segmentation

The audio segmentation of the RWTH system makes use of a log-linear classifier that decides if a time frame corresponds to a segment boundary or not. The features for the log-linear model were chosen as to cover e. g. the variability of the acoustic signal, the speaker homogeneity, and also changes in the acoustic conditions.

In addition, the set of log-linear features was augmented by information obtained by a one-pass recognition on the unsegmented audio data. This includes the number of words within a segment as well as the time stamps of sentence boundary tokens hypothesized by the recognizer.

Finally, an HMM-based classifier was used to detect speech/non-speech, music, and noise segments. From this information, additional features were derived and included into the log-linear segment boundary classifier. Further details on the RWTH segmentation software can be found in [25].

5.7. Recognition setup

Most RWTH systems rely on five subsequent recognition passes, as depicted in Figure 2. In an initial unadapted pass, a first transcription was obtained which formed the basis for the second, CMLLR-adapted recognition pass.

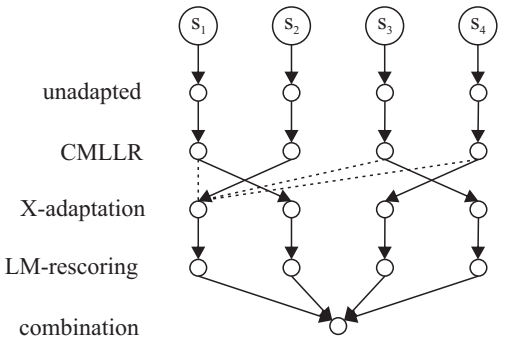

Figure 2: *Schematic view of the RWTH recognition process*

Table 12: *RWTH improvements in terms of WER obtained by different methods, measured on the 2010 development corpus*

	DE (s_1)	*FR (s_1)*
baseline	26.2 %	32.2 %
VTLN	21.9 %	25.5 %
+CMLLR	20.2 %	23.2 %
+MPE	19.9 %	22.6 %
+X-adaptation	19.9 %	22.5 %
+LM-rescoring	19.8 %	22.3 %
+CNC	17.3 %	20.9 %

The resulting transcriptions then were exchanged between subsystems for cross-adaptation, leading to the third full recognition pass.

As for the full recognition only a pruned LM was used, third pass lattices were rescored using the full LM. System combination was applied as a last step. The lattices were converted to a confusion network (CN) by an iterative procedure. Afterwards, the final transcription was obtained by CN combination as presented in [5]. In case of the German system, a post-processing step was also added to concatenate numbers to avoid spelling errors.

5.8. Results

For both languages confusion network system combination (CNC) of several subsystems was used to produce the final output. Table 12 shows the WER for each pass of the single best subsystem and the final CNC result using all subsystems.

6. System Combination

The Quaero SLT evaluation compared performance on manual reference transcriptions to those resulting from a combination of the Quaero STT systems. The single-best output from the three different sites for each language were combined by performing *Recognizer Output Voting Error Reduction* (ROVER) combination [26] on the respective CTM files.

The parameter settings for the combination were determined empirically, by trying out several configurations on the 2010 development set. For German the best performance was obtained by performing a majority vote among the inputs. For French the best performance was obtained by voting based on the maximum confidence scores. The confidence score for NULL transitions was set to the number of arcs in the correspondence set divided by the number of input systems (this method is called *maxconfa* by the NIST ROVER program.) Alignment of the input hypotheses was done taking the timing information of the words into account.

Table 13 shows the case-insensitive and case-sensitive word error rates achieved with the combination on the 2010 Quaero development and evaluation sets.

For German, the best system output in the ROVER combination achieved a case-insensitive WER of 16.9%, and a case-sensitive WER of 21.2% on the 2010 evaluation set. Thus, for the case-insensitive case the ROVER combination did not give any gain, however, in the—for speech translation relevant—case-sensitive case, the WER was lowered by 14.6% relative to 18.1%.

The best French system in the ROVER combination scored a case-insensitive WER of 20.8%, and a case-sensitive WER of 21.9%. Therefore, the ROVER combination reduced the case-insensitive WER by 5.8% relative, and the case-sensitive WER by 17.4% relative to 20.8%.

Table 13: Result of the Rover Combination for German and French; Case-Insensitive (ci) and Case-Sensitive (cs) Word Error Rate

Language	dev2010 [%WER]		eval2010 [%WER]	
	ci	cs	ci	cs
French	18.1	19.2	19.6	20.8
German	17.0	18.0	17.0	18.1

7. Quaero STT evaluations

In 2011, speech to text transcription was assessed for 9 languages respectively. This was the fourth evaluation of three primary Quaero languages (English, French, German), third evaluation for the Russian and Spanish languages, second for Greek and Polish, and baselines for the Italian and Portuguese languages. All partners involved in technology development participated, with a total of 47 submissions. All sites submitted systems for at least 5 languages and there were at least two submissions for all languages except Italian. LNE distributed an evaluation plan as well as scoring tools. The development data consisted of the previous year's eval data for the 7 languages, with 3 hours development data distributed for the two newly introduced languages.

The data are categorized into Broadcast News and more varied data including talk shows, debates, Web podcasts collectively called Broadcast Conversation. Table 14 summarizes the results in

Table 14: *Summary of test results (case-insensitive WER) for the 9 languages, with proportion of Broadcast News (BN) and Broadcast Conversation (BC). Since the first evaluation for the Italian and Portuguese languages was held in 2011, the P3 eval column gives the results on the dev data set.*

	P3 Eval (2010)		*P4 Eval (2011)*	
Language	*BN/BC*	*WER (%)*	*BN/BC*	*WER (%)*
English	50/50	17.3	30/70	19.8
French	50/50	19.0	30/70	14.9
German	50/50	16.9	30/70	17.4
Russian	50/50	19.2	30/70	18.3
Spanish	50/50	13.6	30/70	15.9
Greek	70/30	20.7	30/70	16.9
Polish	70/30	20.0	30/70	12.5
Italian	50/50	22.8*	50/50	18.0
Portuguese	50/50	28.5*	50/50	22.7

terms of case-insenstive WER for the 2011 STT evaluation. Case-sensitive scoring results in an absolute increase in WER of about 1%.

8. Discussion

There has been steady progress in reducing the word error rates from year to year. The relative word error rate was reduced by over 15% yearly for the three primary languages (English, French and German) for the first three years, and by as much 25% relative for some languages. It is not easy to compare the word error rates on the 2010 and 2011 data, since the latter contain a much larger proportion of conversational speech.

Other ongoing work in Quaero addresses automatic punctuation and the role of prosody, as well as a comparison of recognition errors by both machines and humans in an effort to determine if the errors are due to inherent ambiguity in the speech signal or to inaccurate modeling. While in this evaluation the speech transcripts were provided to the SLT systems, more closely coupled interfaces (confusion networks, confidence scores, post-processing for number and entity conversion) are being investigated by some of the partners.

9. Acknowledgments

This work has been partially financed by OSEO under the Quaero program. The authors would also like to thank Elena Lopez, Rita Sidabraite, Eric Bilinski and Olivier Galibert without whom the evaluations would not have been possible.

10. References

[1] K. Boudahmane, B. Buschbeck, E. Cho, J, Crego, M. Freitag, T. Lavergne, H. Ney, J. Niehues, S. Peitz, J. Senellart, A. Sokolov, A. Waibel, T. Wandmacher, J. Wuebker and F. Yvon, "Advances on Spoken Language Translation in the Quaero Program", *International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, Dec. 2011.

[2] L.F. Lamel, J.L. Gauvain, and M. Eskénazi, "BREF, a large vocabulary spoken corpus for French," *ESCA EuroSpeech'91*, pp. 505-508, Genoa, September 1991.

[3] H. Soltau, F. Metze, C. Fügen and A. Waibel. "A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment," *IEEE ASRU*, Madonna di Campiglio, Italy, December 2001.

[4] M. J. F. Gales. "Semi-tied covariance matrices," *IEEE ICASSP*, **2**:657-660, Seattle, 1998.

[5] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination", *NIST Speech Transcription Workshop*, 2000.

[6] T. Schultz, "GlobalPhone: A Multilingual Speech and Text Database Developed at Karlsruhe University," *ICSLP'02*, pp. 345-348, Denver, Sept. 2002.

[7] T. Schultz, A.W. Black, S. Badaskar, M. Hornyak, and J. Kominek, "Spice: Web-based tools for rapid language adaptation in speech processing systems", *Interspeech'07*, pp. 2125-2128, Antwerp, Belgium, August 2007.

[8] T. Vu, T. Schlippe, F. Kraus, and T. Schultz, "Rapid Bootstrapping of five Eastern European Languages using the Rapid Language Adaptation Toolkit", *Interspeech'10*, pp. 865-868, Makuhari, Japan, September 2010.

[9] A. Stolcke, "SRILM - An extensible language modeling toolkit", *IEEE ICASSP'02*, **2**:901-904, Orlando, FL, May 2002.

[10] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion", *Speech Communication*, **50**(5):434-451, May 2008.

[11] J.L Gauvain, L. Lamel and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication,* **37**(1-2):89-108, 2002.

[12] L. Lamel and J.L. Gauvain, "Speech processing for audio indexing," *GoTAL 2008 - Advances in NLP, no.5221/2008 LNCS,* pp. 4-15. Springer Verlag, 2008.

[13] J.L. Gauvain, L. Lamel, and G. Adda, "Partitioning and transcription of broadcast news data," *ICSLP'88*, Sydney, Australia, December 1998, pp. 1335-1338.

[14] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, **9**(2), pp. 171-185, 1995.

[15] H. Hermansky, "Perceptual linear prediction (plp) analysis for speech," *J. Acoust. Soc. Amer.*, **87**:1738-1752, 1990.

[16] P. Fousek, L. Lamel and J.L. Gauvain, On the Use of MLP Features for Broadcast News Transcription, *TSD'08*. LNCS 5246/2008, 303.10, Springer Verlag, Berlin/Heidelberg, 2008.

[17] P. Schwarz, P. Matějka and J. Černocky, "Towards Lower Error Rates In Phoneme Recognition," *TSD'04*, pp. 465-472, Brno, September 2004.

[18] F. Grézl and P. Fousek, Optimizing Bottle-Neck Features for LVCSR, *ICASSP'08*, pp. 4729-4732, Las Vegas, 2008.

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

[20] H. Schwenk, *Continuous Space Language Models*, Computer, Speech & Language, **21**:492-518, 2007.

[21] H. Schwenk and J.L. Gauvain, Training Neural Network Language Models On Very Large Corpora, *JHLT/EMNLP*, pp. 201-208, Vancouver, 2005.

[22] C. Plahl, R. Schlüter and H. Ney, "Hierachical Bottle Neck Features for LVSCR", Interspeech'10, Makuhari, Japan, pp. 1197-1200, October 2010.

[23] D. Giuliani, G. Stemmer and F. Brugnara, "Adaptive training using simple target models", *IEEE ICASSP'05*, f1:997-1000, Philadelphia, PA Mar. 2005.

[24] D. Povey and P.C. Woodland, "Minimum Phone Error and I-smoothing for Improved Discriminative Training", *IEEE ICASSP'02*, **1**:105-108, Orlando, FL, May 2002.

[25] D. Rybach, C. Gollan, R. Schlüter and H. Ney, "Audio Segmentation for Speech Recognition using Segment Features", *IEEE ICASSP'09*, pp. 4197-4200, Taipei, Taiwan, April 2009.

[26] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *IEEE ASRU'97*, pp. 347-354. Santa Barbara, CA, December 1997.