Development of a Speech-to-Text Transcription System for Finnish*

Lori Lamel[1] *and Bianca Vieru*[2]

[1]Spoken Language Processing Group
CNRS-LIMSI, BP 133
91403 Orsay cedex, France
lamel@limsi.fr

[2]Vecsys Research
3, rue Jean Rostand
91400 Orsay, France
vieru@vecsysresearch.com

Abstract

This paper describes the development of a speech-to-text transcription system for the Finnish language. Finnish is a Finno-Ugric language spoken by about 6 million of people living in Finland, but also by some minorities in Sweden, Norway, Russia and Estonia. System development was carried out without any detailed manual transcriptions, relying instead on several sources of audio and textual data were found on the web. Some of the audio sources were associated with approximate (and usually partial) texts, which were used to provide estimates of system performance.

1 Introduction

Traditionally speech-to-text transcription (STT) systems are trained on large amounts of carefully transcribed speech data and huge quantities of written texts. However obtaining the needed transcribed audio data remains quite costly and requires substantial supervision. Several research directions have addressed reducing these costs [6] and much of the recent audio training data, as in the DARPA Gale program are associated with quick transcriptions (QTR) [7]. For certain audio sources, it is possible to find associated texts, ranging from quite accurate, but usually incomplete, transcriptions, to closed captions, summaries or other less closely related texts. A variety of approaches have been investigated most relying on supervision from a language model. The approaches differ in their details: use or not of confidence factors [8] or [9], [10], doubling vs iterative training [11] and the amount of data used.

In this study, system development is also lightly supervised, in that no detailed annotations are available for the development and test data. Initially approximate reference transcriptions were used to assess both acoustic and language models during system development. Only afterward were the transcripts manually corrected in order to have a better estimate of the true performance.

The next section gives an overview of the characteristics of the Finnish language, followed by a description of the approach and corpus used in this study. This is followed by a description of the language models, phone set and acoustic models, after which experimental results are provided.

2 Finnish Language

Part of Uralic languages, Finnish is a Finno-Ugric language spoken by about 6 million of people living in Finland, but also by some minorities in Sweden, Norway, Russia and Estonia.

Finnish shares a basic vocabulary with the other Uralic languages and has various derivational suffixes. It has regular letter-to-sound correspondences, which simplifies the problem of pronunciation modeling. While Finnish has a smaller core vocabulary than English, it allows creation of new words by extensive use of agglutination, resulting in a very large lexical variety.

Most of the reported speech-to-text transcription results for the Finnish language are substantially worse than results reported for more resourced languages such as English or French. A first explanation could be that the extensive use of agglutination in Finnish which has impact on the language modeling difficulties. In [1] it is highlighted that using a 20K word vocabulary in English gives a lower OOV rate than a 500000-word vocabulary in Finnish. For example 40-million-word English corpus contains about 190000 distinct words, while the corresponding Finnish corpus contains about 1.9 million unique words. A proposed solution to this problem is the decomposition of words into morphs as shows in [2, 3].

But another explanation of this poor results is the lack of suitable speech and text training data resources. If in 2002, about 72% of the websites were in English although that

*This work was supported in part by OSEO under the Quaero program and by the European Commission under the VITAL project.

was the language of only a third of the Web users, in 2003 Finnish is found to be used in 1% of a random selection of web pages [4].

The results are less dramatic when one looks at the languages used by people to communicate with each other via the Web: they then prefer to use their mother tongue. Moreover, the recent rise of the blog, reflecting a desire to reach a smaller audience closer to the writer, could allow an increase in the range of languages used on the Internet, in particular French [5].

3 Approach and Corpus

The general approach taken in the work is similar to that of [11, 12, 13] in that a speech recognizer is used to provide "approximate" transcripts for acoustic model training. The audio data is transcribed in batches, and in successive iterations the models are trained on more data. In [14] an analysis of training behavior is compared for supervised and unsupervised approaches.

In contrast to previous studies where audio and text data were available for model training, the first challenge in this study was locating audio and text data in Finnish. Three types of audio data were found. The first data are from a website which we refer to as BN Learning website, diffusing news audio data with close transcriptions targeting an audience of non-native speakers of Finnish. The data on this site use a simplified language so as to be accessible to foreigners. A total of 31 hours of audio with corresponding approximate transcriptions (102k words) have been downloaded since November 2007. A second data set containing 19 hours of audio with approximate transcriptions was downloaded from the Finnish News Agency. These audio correspond to short newswires diffused hourly for native Finnish speakers. The transcripts cover only part of the audio and are not aligned.

Since the initial word error rate estimates were quite low on these data compared to previously published results for Finnish, it was decided to extend the range of data sources and types (general news, special reports, interactive shows).[1] A total of 190 hours of varied broadcast data were collected from a variety of Finnish sources. The audio data used in this study are summarized in Table 3. In addition to the audio data and (when available) associated transcripts, 30M words from text materials were collected.

Initial acoustic and language models were built using just the BN learning corpus. Then the FNA data were added, and finally some of the more general data. In order to provide supervision in acoustic model training, the language models used in the early decoding stages of the audio data were heavily biased, being trained on texts from the same epoch of the BNL transcripts. The language models used for test purposes were initially also only trained on the BNL data, but quickly additional texts were included. It should be noted that both AM and LM development were ongoing, as the text normalization was progressively improved.

Texts	Transcription		Newspapers
	BNL	FNA	
Train	78K	193K	30M
Dev	24K	48K	

Table 1: Text corpora used for language modeling.

4 Language models

Texts from over 20 different sources, mainly newspapers, formed the language model training corpus. As can be seen in Table 1 approximate transcriptions of audio data represent less than 1% of the text corpus. Concerning the transcriptions, it should be noted that the BN learning texts uses a substantially simplified language compared to standard Finnish broadcast news.

In Finland, newspaper articles are written in Finnish or in Swedish, both being official languages. Also sometimes small citations or entire articles can be found in English, Russian, Estonian or other languages in texts downloaded from Internet. Since the language is not always clearly indicated in the texts, text based language identification using the program TextCat [15] was run on each processed paragraph and only Finnish paragraphs were retained.

As is standard practice, the texts were split into sentences and the main punctuation was removed. During normalization all words were converted to lowercase, and words with a dash or a colon were separated, keeping the dash and colon as words. Numbers were transformed to a full, spoken form. This is quite complicated for the Finnish language which has 15 declensions cases and all parts of numbers should be declined. Some cases are constructed by adding suffixes, such as 's' in ordinals, after each component number. Given the complexity of expanding numbers into words for different cases, and our lack of knowledge about the Finnish language, it was decided to first only use the nominative case. After processing, the texts contained a total number of about 30M words, with a vocabulary size of about 1.4M words.

Finnish is an agglutinative language, using suffixes to ex-

[1] These sources were found by a native Finnish speaker who also, after we had developed a system with lightly supervised references, corrected the transcripts of the BN Learning and FNA news data.

press grammatical relations and also to derive new words, so the vocabulary expands rapidly. There is no grammatical gender for nouns however all 15 cases are even even for proper names. This makes the loan words difficult to treat since each word could appear in each of these 15 cases. For example, the following forms were found in the texts:

> ***Bush*** *Bushia Bushien Bushiin Bushilla Bushille Bushilta Bushin Bushissa Bushista Bushit Bushkin*
>
> ***Obama*** *Obamaa Obamaan Obamalla Obamalle Obamalta Obaman Obamassa Obamasta Obamat*

A list of words was selected by interpolation of unigram models trained the normalized texts from different newspapers and the approximate transcripts associated with the audio data. The Morfessor [16] decompounding algorithm is applied to this list to determine possible word decompositions. For example, for the word *elinaikakerroin* (survival factor) Morfessor proposes *elin* + *aika* + *kerroin*, which is mapped to *elin_ _aika_ _kerroin* in order to keep track that the lexical entries result from a decomposition. In order to avoid creating too many small, easily confusable lexical entries, a minimum of 3 characters per unit was imposed. All of the texts are decomposed using the selected decompositions proposed by Morfessor. Since the resulting lexical entries differentiate words from the decomposed forms, the language models decide the appropriate form and the forms in the hypotheses can simply be glued back together. The total number of tokens in the text corpus is increased as a result of word decomposition, but the number of distinct word forms is divided by two.

As mentioned in Section 3 biased n-gram language models were constructed to decode the audio by training on only the associated approximate transcriptions collected from the same period (usually 1 month) in order to provide strong, but flexible supervision. These initial LMs were based on full word lexical entries (no decomposition) and were used only for the first acoustic models.

For the second iteration, language models trained on all the transcripts from the same year and type as the audio data were constructed in order to have a more general LM. The LMs were interpolated with a general language model trained on the entire text corpus, with each component LM having an equal mixture weight.

Different language models were used in speech recognition experiments. For most of the experiments, the language models use a 300k word list optimized on the BNL+FNA dev data. The n-gram language models were obtained by interpolation of backoff n-gram language models trained on separate subsets of the available language model training texts using the modified Kneser-Ney smoothing. The characteristics of the 300k 4-gram language models are summarized in Table 2. The mixture weights were automatically chosen using the EM algorithm to minimize the perplexity of the development data. It can be seen that the perplexity and OOV rates of the BNL data and the FNA dev are much lower than the test data.

Type	BNL_dev	FNA_dev	FNA_test	BN_test
OOV	0.67	1.81	4.01	3.85
ppx	193	386	2418	2668

Table 2: Perplexity (PPX) and Out Of Vocabulary (OOV) rates for the different sets of dev and test data using a 300k LM. The LM mixture weights were tuned on the dev data.

5 Phone Set & Acoustic models

Words of foreign origin excluded, Finnish is written with 8 letters for vowels and 13 for consonants. All the vowels and almost all the consonants can be either short or long sounds. The phone set used in this work is composed of 42 phones: 16 vowels, 27 consonants and three units for silence, breath and filler. The long and short phones are represented with separate symbols and have separate acoustic models. Standard Finnish is basically a phonetic language where each letter corresponds to one and the same phoneme, and each phoneme corresponds to one and the same letter [17]. So, with very few exceptions, the lexicon observes a strict correspondence between letters and phonemes, with a low number of variants (avg 1.1 pronunciations/word).

A multi-language, cross-language bootstrapping [18] was used to initialize the acoustic models. Phones from English, French, German, Italian and Arabic were mapped to Finnish phones, and models extracted from corresponding acoustic model sets served as initial seed models. The first month of BN learning data was decoded using these models and a language model built only on transcriptions of that month (with a 22k word LM). The acoustic models were trained in a lightly supervised manner [13], one month at a time until the full 14 hours of speech from the BN Learning (BNL) corpus was used. For the first stages a 22k LM was used to decode the audio data. Data from the standard BN (FNA) were then progressively added with larger models trained after each step.

The standard cepstral features (perceptual linear prediction - PLP) were used. The PLP feature vector has 39 cepstral parameters: 12 cepstrum coefficients and the log energy, along with the first and second derivatives. The acoustic models are tied-state, left-to-right context-dependent, HMMs with Gaussian mixtures. The triphone-based context-dependent phone models are word-independent, but

Audio	**Learning**	**FNA**	**BN**
Train	19h	11h	170
Dev	7h	4h	-
Test	-	4h	16h

Table 3: Audio corpus (in hours) used for training, dev and test for the Finnish STT system.

Model			Audio corpus	
set	ctx	#Gaussians	Hours	Sources
BN0	8345	190k	26	BNL+FNA
BN1	9713	239k	35	+BN 9 hrs
BN2	10568	272k	42	+BN 16 hrs
BN3	12493	355k	63	+BN 37 hrs
BN4	18268	369k	195	+BN 169 hrs

Table 4: Characteristics of different acoustic model sets.

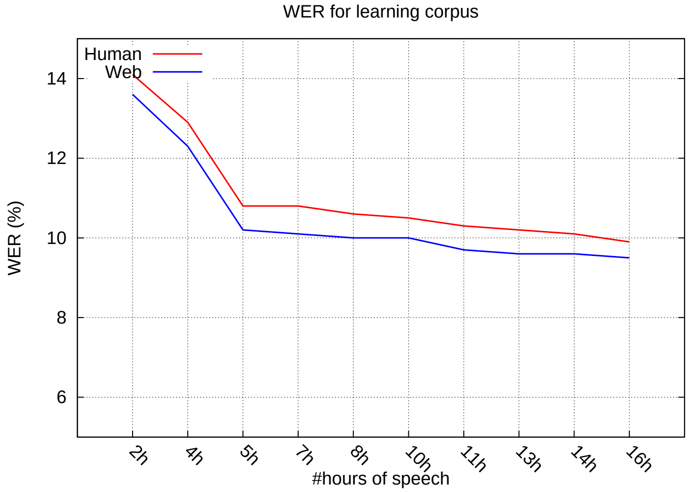

Figure 1: Performance on BN Learning development data using a 700k LM estimated on a 10M word text corpus.

word position-dependent. The tied states are obtained by means of a decision tree. The acoustic models are gender-independent and speaker-adaptive trained (SAT). Silence is modeled by a single state with 1024 Gaussians. The best model trained on only the BN Learning corpus cover about 5.6k phone contexts, with 3.7k tied states and 32 Gaussians per state. With the additional 11 hours of FNA data, the acoustic models cover 8k contexts and 6k tied states. These models, trained on the pooled data were also then MAP [19] adapted to each audio corpus. As more of the varied BN data was progressively added, larger models were built, with the largest covering about 18k contexts as shown in Table 4.

6 Experimental results

This section reports a series of experiments assessing recognition performance as a function of the available acoustic and language model training data. The system is based on the LIMSI broadcast news transcription [20] was used. It has two main components, the audio partitioner and the word recognizer. During development of the Finnish STT system, all evaluation was done using selected portions of the web transcriptions as references (based on string alignments). These may be inexact and often contain either fewer or more words than in exact transcriptions. After system development, a native Finnish speaker corrected these transcriptions and a real scoring was realized.

Figure 1 shows the recognition results using web transcripts and the corrections made by a native Finnish on the BN learning corpus. In these experiments, acoustic models were built using only the BNL data with a vocabulary of 700k decomposed words and language models built on a 8M text corpus available at the time. As can be seen in the figure, the two error curves closely follow each other, with slightly optimistic results with the approximate transcripts.

As in [12] a speech recognizer was used to automatically transcribe unannotated data and generating ”approximately” labeled training data. As the amount of training data increases iteratively, more accurate acoustic models are obtained, which can then be used to transcribe another set of unannotated data. The data were added progressively, choosing the data with good likelihood scores first [8, 9]. The characteristics of the acoustic models are given in Table 4.

Figure 2 shows that using web references for scoring can give an idea of system performance of different acoustic sets. For each set of curves, the solid line corresponds to scoring with approximate web transcripts and the dotted lines scoring with manually corrected references when available. It can be seen that although the absolute levels are different, the behavior of the curves are quite similar. There is a particularly a big difference for the FNA_test, which is due to the fact that available web transcriptions do not cover all of the audio data, so the insertion rate is very high. In contrast, the curves are very close for the BNL dev data for which close approximate transcriptions are available. It can also be seen that as progressively more varied BN data are included in the training, the BNL and FNA results slowly degrade. The first set of models (BN0) are trained on only FNA and BNL data, so these are closer to the dev and test data. These experiments all used the same 300k word list

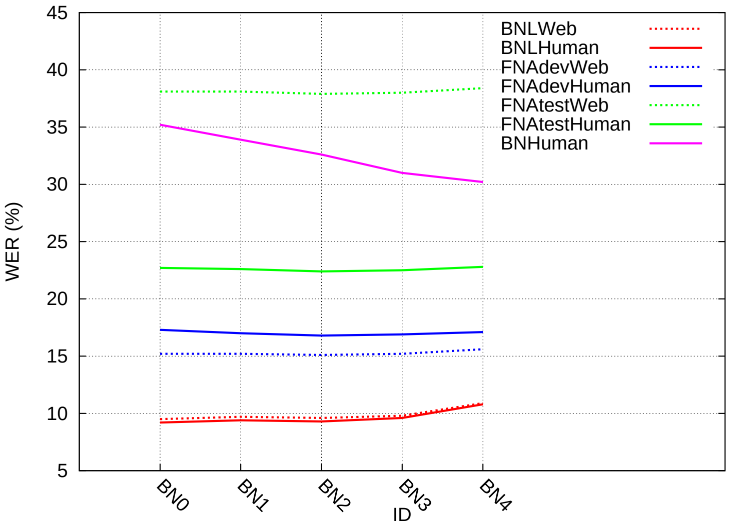

Figure 2: System performance with manual or web references (when available). Acoustic models are built on BNL and FNA, and progressively more BN data. A 300k vocabulary obtained by interpolation of 1-grams on BNL+FNA dev was used.

	BNL	FNA_dev	FNA_test	BN
Web	8.8	14.4	21.9	-
Human	9.1	16.3	37.3	29.4

Table 5: WER with manual references of best system for each type of data with a two pass decoding and unsupervised acoustic model adaptation.

selected by interpolating 1-grams so as to optimize the coverage of the BNL and FNA dev data, and language models trained on the 34M word (decompound words) corpus.

Table 5 gives the best results obtained on different data types. These results are obtained using a 2 pass system, with unsupervised acoustic model adaptation between decoding passes [20]. The acoustic models are also specific to each data type, being MAP [19] with the available audio training data from each audio corpus (using the automatic transcripts).

7 Conclusions

This paper has described the development of a speech-to-text transcription system for the Finnish language. The first task was locating appropriate resources for acoustic and language model training, and system assessment. In doing so the methodology used in lightly supervised or unsupervised acoustic model training has been extended to system development since no carefully transcribed development data was available for model optimization. Transcription word error rates were reported with approximate web transcripts that were used during system development and with manual transcripts that were later created, and although the approximate transcripts give an optimistic estimate of the true word error rates they were found to be useful for system optimization.

References

[1] Teemu Hirsimki, *Advances in unlimited-vocabulary speech recognition for morphologically rich languages*, Phd dissertation, Helsinki University of Technology, Department of Information and Computer Science, 2009.

[2] V. Siivola, T. Hirsimäki, M. Creutz, and M. Kurimo, “Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner,” in *in Proc. Eurospeech*, 2003, vol. 20, pp. 2293–2296.

[3] T. Hirsimaki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkonen, “Unlimited vocabulary speech recognition with morph language models applied to finnish,”*Computer Speech & Language*, vol. 20, no. 4, pp. 515–541, October 2006.

[4] Edward T. O’Neill, Brian F. Lavoie, and Rick Bennett, “Trends in the evolution of the public web, 1998 - 2002,” *D-Lib Magazine*, vol. 9, 2003.

[5] Anne de Beer and Grard Blanc, “La diversit des langues sur internet,” *Futuribles*, vol. 329, pp. 29–36, 2007.

[6] O. Kimball, C.L. Kao, R. Iyer, T. Arvizo, and J. Makhoul, “Using quick transcriptions to improve conversational speech models,” *INTERSPEECH*, pp. 2265–2268, 2004.

[7] C. Cieri, D. Miller, and K. Walker, “The fisher corpus: a resource for the next generations of speech-to-text,” *LREC*, pp. 69–71, 2004.

[8] C. Gollan, M. Bisani, S. Kanthak, R. Schlüter, and H. Ney, “Cross domain automatic transcription on the TC-STAR EPPS corpus,” *ICASSP*, vol. 1, pp. 825–828, 2005.

[9] M. Bisani and H. Ney, “Joint-sequence models for grapheme-to-phoneme conversion,” *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[10] F. Wessel and H. Ney, “Unsupervised training of acoustic models for large vocabulary continuous speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 23–31, 2005.

[11] J. Ma and R. Schwartz, “Unsupervised versus supervised training of acoustic models,” *INTERSPEECH*, pp. 2374–2377, 2008.

[12] L. Lamel, J.L. Gauvain, and G. Adda, “Lightly supervised acoustic model training,” *ITRW ASR*, vol. 1, pp. 150–154, 2000.

[13] L. Lamel, J.L. Gauvain, and G. Adda, “Lightly supervised and unsupervised acoustic model training,” *Computer Speech and Language*, vol. 16, pp. 115–129, 2002.

[14] S. Novotney and R. Schwartz, “Analysis of low-resource acoustic model self-training,” *INTERSPEECH*, pp. 244–247, 2009.

[15] G. van Noord, “http://www.let.rug.nl/vannoord /textcat/,” .

[16] M. Creutz and K. Lagus, “Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0,” Tech. Rep., Helsinki University of Technology, 2005.

[17] F. Karlsson, *Finnish: An Essential Grammar (Routledge Grammars)*, Routledge, 1st edition, 1999.

[18] J. Loof, C. Gollan, and H. Ney, “Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a polish speech recognition system,” *INTERSPEECH*, pp. 88–91, 2009.

[19] 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, pp. 291–298, 1994.

[20] J.L. Gauvain, L. Lamel, and G. Adda, “The limsi broadcast news transcription system,” *Speech Communication*, vol. 37, pp. 89–108, 2002.