Computer-assisted closed-captioning of live TV broadcasts in French

G. Boulianne, J.-F. Beaumont, M. Boisvert, J. Brousseau, P. Cardinal,
C. Chapdelaine, M. Comeau, P. Ouellet, F. Osterrath

Centre de recherche informatique de Montréal (CRIM)
Montréal, Canada
Gilles.Boulianne@crim.ca

Abstract

Growing needs for French closed-captioning of live TV broadcasts in Canada cannot be met only with stenography-based technology because of a chronic shortage of skilled stenographers. Using speech recognition for live closed-captioning, however, requires several specific problems to be solved, such as the need for low-latency real-time recognition, remote operation, automated model updates, and collaborative work. In this paper we describe our solutions to these problems and the implementation of a live captioning system based on the CRIM speech recognizer. We report results from field deployment in several projects. The oldest in operation has been broadcasting real-time closed-captions for more than 2 years.

Index Terms: speech recognition, closed-captioning, model adaptation.

1. Introduction

While closed-captioning of TV programs becomes increasingly available in English (about 90% of televised contents), hardly 60% of French broadcast news are closed-captioned. For live interviews or reports, the percentage is lower still. This restricted accessibility of information to French speaking deaf and hearing impaired viewers is in large part due to a lack of available technologies. The federal government agency that oversees Canadian TV (CRTC) is aware of the situation and has begun to take action by compelling Canadian broadcasters to improve on the quantity and quality of their closed-captioning, particularly for live broadcasts.

In this context a first prototype was produced in a joint project involving the GTVA Network and CRIM's speech recognition team to adapt CRIM's transducer-based large vocabulary French speech recognizer. Trial broadcasts started in 2003 and live news captions are broadcast on a regular basis since February 2004. Since then, our system has been evaluated in trials for the captioning of Canada's House of Commons parliamentary debates, and is currently producing live captioning of RDS (national sports network) NHL Saturday night hockey games.

Shadow speakers, who listen to the original audio, interpret it and repeat it to the system, circumvent the problems of difficult acoustics and speaker variability. Even then, reaching acceptable accuracy under low-latency constraints and evolving news topics remains a challenging problem for current speech recognition technology. We will first describe the system architecture and initial recognition setup, then the methods we developed to maintain and enhance initial performance for several years through automated vocabulary, language and acoustic model updates. Finally we will report results on 3 ongoing live captioning projects.

2. Architecture

The overall closed-captioning process proceeds along the following steps. Audio is sourced from a newscaster to a shadow speaker who repeats the spoken content. The respoken audio is then sent over a computer network to a speech recognizer which produces transcriptions that are filtered, formatted, and fed to the broadcaster's closed caption encoder.

The architecture (Figure 1) was designed to allow several shadow speakers to collaborate during a live session, through a lightweight user interface running on each speaker workstation, a shared database, a speech recognition server and encoder servers which dispatch captions to Line 21 encoders, all communicating through a TCP/IP based protocol. System components can be physically located anywhere on the network, which facilitates captioning of remote sites.

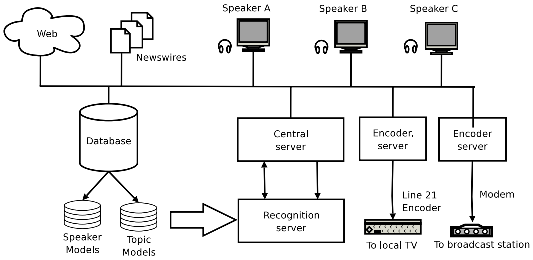


Figure 1: Closed-captioning system architecture.

2.1. Server side

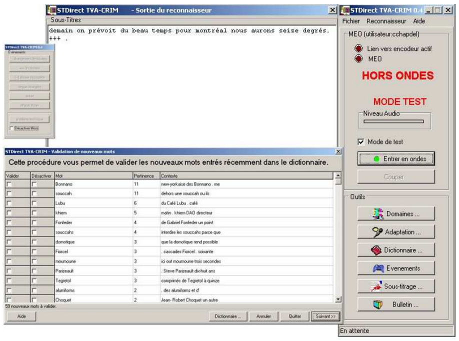
The database is the central repository for words and their pronunciations, and tracks the dynamic status of words, their association with topics, and their origin. It is also used for administrative tasks such as user profile management and logins.

A captioning "configuration" is defined as a quadruple $C_{TR} = \{T, V_T, G_T, A_R\}$: a topic T, the set of words V_T currently active for this topic (vocabulary), the language model G_T associated with this topic, and an acoustic model A_R for a speaker R. Before a live captioning session, a number of configurations are preloaded in memory, so that switching between topics and speakers happens instantly during the session.

The recognition server handles all speech related tasks such as recognition, acoustic model adaptation, grapheme-to-phoneme conversion, and storage of recordings. The CRIM recognition engine uses a precompiled, fully context dependent finite-state transducer [1] and runs in a single pass at 0.7 times real-time (on average) on a Pentium Xeon 2.8 GHz CPU. There is a minimum delay of 1 second and maximum delay of 2 seconds between the input shadow speech and the text output.

2.2. Client side

The user interface (Figure 2) runs on each shadow speaker workstation and provides user main functionalities: pre-production, production and post-production.



STDirect TVA-CRIM - Sortie du reconnaisseur

Sous-Titres

```
demain on prévoit du beau temps pour montréal nous aurons seize degrés.
+++ .
```

STDirect TVA-CRIM - Validation de nouveaux mots

Cette procédure vous permet de valider les nouveaux mots entrés récemment dans le dictionnaire.

Valider	Désactiver	Mot	Pertinence	Contexte
□	□	Bonnano	11	new-yorkaise des Bonnano . me
□	□	souccah	11	dehors une souccah ou ils
□	□	Lubu	6	du Café Lubu . café
□	□	khiem	5	matin . khiem DAO directeur
□	□	Fonteder	4	de Gabriel Fonteder un point
□	□	souccahs	4	interdire les souccahs parce que
□	□	domotique	3	que la domotique rend possible
□	□	Fiorcel	3	. cascades Fiorcel . soixante
□	□	moumoune	3	ici oui moumoune trois secondes
□	□	Parizeault	3	. Steve Parizeault dix-huit ans
□	□	Tegretol	3	comprimés de Tegretol à quinze
□	□	alumiforms	2	. des alumiforms et d'
□	□	Choquet	2	Jean- Robert Choquet un autre

59 nouveaux mots à valider.

STDirect TVA-CRIM 0.4

MEO (utilisateur:cchapdel)

● Lien vers encodeur actif
● MEO

HORS ONDES

MODE TEST

☑ Mode de test

En attente

Figure 2: Shadow speaker workstation client software.

In the pre-production stage, shadow speakers use the dictionary editor interface to verify word pronunciations, check word association with topics, add new words or validate new entries proposed daily by the system. They also select an encoder configuration (which will establish a modem connection if needed).

During live production, shadow speakers listen and repeat through a head-mounted headphone/microphone combination connected to the workstation. They change topics, insert punctuation, indicate speaker turns and insert other symbols, while speaking, by using a video-game control pad.

After the session ends, they use a correction interface to listen to their recorded voice and correct any recognition errors. This data is stored to be used later for supervised adaptation of acoustic and language models.

3. System initialization

In this section we describe the baseline models that served as the starting point of the French broadcast news system. Section 4 will then describe how the baseline system is adapted every night to yield the actual production system used in everyday operations.

3.1. Acoustic models

Baseline acoustic models are speaker-independent, gender-dependent, continuous HMM models, with 5535 cross-word triphone models sharing 1606 state output distributions, each a mixture model of 16 Gaussians with individual diagonal covariances. Input parameters are 39-dimensional vectors of 13 static MFCC plus energy with first- and second- order derivatives. The models were trained on the *Transtalk* database, which contains 40 hours of clean speech read by 30 French Canadian speakers.

3.2. Language model

The starting point for the language model is a collection of French Canadian newspapers, news report collected from the Web, and broadcaster news archives which provides a total of 175 million words of text. We describe here the procedure used to estimate language models for the news domain. Other domains, such as parliamentary debates and hockey games, follow the same procedure but use fewer subtopics.

News texts are automatically classified into 8 pre-defined topics: *culture, economy, world, national, regional, sports, weather, traffic* using a Naive Bayes classifier [2] trained on topic-labelled newspapers articles. For each topic, larger text sources are also ordered chronologically to be partitioned into older and newer data sets.

A different mixed-case vocabulary is selected for each topic using counts weighted by source size [3]. For each vocabulary, pronunciations are derived using a rule-based phonetizer augmented with a set of exceptions added by hand. The phoneme inventory contains 37 French phonemes and 6 English phonemes.

Using topic-specific vocabularies, a 3-gram language model is estimated for each time/topic partition. Another 3-gram generic language model is also estimated by merging all topics together. The most recent 1% of the texts is withheld from training, and topic- and time- dependent and generic language models are interpolated together to optimize perplexity of this data, to yield 8 topic-specific language models. Finally, language models are pruned to a reasonable size using entropy-based pruning [4].

For weather and traffic, no adequate written source could be found to train a language model. We resorted to synthetic texts produced from a grammar generalized from a few examples, and manual transcriptions. Using even such a small amount of in-domain data resulted in better language models than using a more general topic model such as *regional*.

4. System update

The topics, words, vocabularies, and pronunciations from the baseline system are entered into the database to form the initial captioning system. From that point onwards, the database contents will evolve in time due to automatic updates and manual corrections, and all recognition models will be derived from the content of the database.

4.1. Unsupervised vocabulary adaptation

In the news, words are introduced everyday. Names such as *Katrina* or *tsunami* appear suddenly while others fade out of frequent usage. Yet, out-of-vocabulary (OOV) words are an important component of the word error rate. Substituting a place or person's name has a worse effect on understanding than using the wrong number or gender agreement (another frequent source of errors in French). Thus it is important that new words are automatically added to each topic vocabulary every day. This adaptation is unsupervised, in the sense that word-topic associations are not given a priori, but must be estimated by the adaptation procedure itself.

The system uses a configurable Web crawler that extracts texts from a number of Web sites, newswire feeds and the broadcaster's internal archives. These texts are stored in an XML database together with meta-data collected by the crawler. The topic classifier (section 3.2) assigns to each text a probability for each topic.

Each day, newly collected texts are compared against the current topic-specific vocabularies (captions corrected in post-production are also considered a source of text). Potential new words are passed through a series of garbage filters and are automatically accepted, automatically rejected, or accepted temporar-

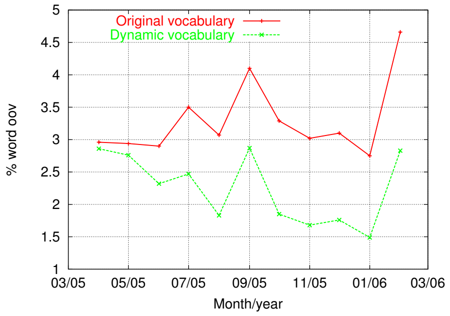


Figure 3: Static and dynamic out-of-vocabulary rates.

ily but proposed to the user for acceptance.

The crawler retrieves about 1 million words of text every night. On average 6000 of these are unknown to the system, but only 200 or so will survive the garbage filters and be added to the database and proposed to the user for verification.

Associations between words and topics have a limited lifetime, so words become inactive in a topic after 60 days; words from the initial baseline vocabularies never become inactive. In this way vocabularies never grow too large or too small.

Figure 3 illustrates the evolution, over almost a year, of the out-of-vocabulary rate for a 20K word topic-independent vocabulary, relative to the reference texts (corrected captions). The full line shows the static vocabulary OOV rate obtained for the unchanging initial vocabulary; the dashed line shows the dynamic vocabulary OOV rate obtained when the vocabulary is updated every day. The vocabulary update is effective, allowing the dynamic vocabulary to produce only around half the static vocabulary OOV rate towards the end of the period.

4.2. Unsupervised language model adaptation

Language models must be adapted in two ways. First, new words must be provided with an adequate language model probability even if they have not been seen in training. Second, all probabilities in the language model, including higher-order n-grams, should be adapted to reflect changes in word usage. Both of these adaptations can only use a small amount of text collected in the last few days, or no data at all in the case of new words added directly by the user. In both cases, adaptation is unsupervised, in the sense that adaptation text has to be classified into topics automatically by the Web crawler.

We interpolate the background (existing) language model unigrams with a unigram model estimated from the adaptation data (before interpolation, in each unigram model, words in the vocabulary that were not observed in training are assigned the same probability as the least frequently observed word in the model). Then higher-order n-gram probabilities in the background model are adjusted using minimum discriminant estimation (MDE), which finds an adapted language model that is as close as possible (in the Kullback-Leibler sense) to the background model, and has the adaptation unigram probabilities as its marginal distribution [5]. This procedure has been found provide good recognition of words added with very small amounts of context (such a the list of hockey players in a team that was not part of the training set) while not degrading the accuracy for already well-trained words.

4.3. Acoustic model incremental adaptation

Acoustic models are subject to gradual performance degradation over long time periods, and changes in the acoustic environment (noise, wall reverberation), voice or microphone positioning also affect the recognition performance. To counter these effects, we use both short-term and long-term adaptation of the speaker-dependent models.

Short-term adaptation (also called “session adaptation”) is done at the beginning of each captioning session. A short news extract is played and repeated by the shadow speaker so that an MLLR transform [6] can be estimated. This procedure does not substantially affect average accuracy, but it reduces variations in error rate across captioning sessions.

Long term adaptation is performed every night. Audio and text from the day’s production of each shadow speaker are aligned using its current model. If enough aligned data has accumulated for a speaker since its model was last updated (typically 40 minutes are required), its model undergoes an MLLR transformation [6] followed by a MAP adaptation [7]. A small part of the data (5 minutes) is held out. If its likelihood is improved, the adapted model becomes the current model. If there is no update, the data is kept available for subsequent training. Using a MAP adaptation scheme guarantees that over long periods, the adapted model will asymptotically converge to the maximum likelihood model. At enrollment time, new shadow speakers start with a copy of the gender-dependent model.

5. Results

In this section we summarize results obtained over a period of a few years, while producing closed-captions in live trials in the course of three research projects.

The first project goal was closed-captioning live parts of the three daily news shows of GTVA, the largest North-American French private network. This is the most complex task, requiring the 8 topics mentioned in section 3.2. Captioning is done on-site, with GTVA trained personnel, since February 2004.

The second project is Canada’s House of Commons (HoC) parliamentary debates captioning. It is a more restricted domain, but captions must be verbatim. In addition, shadow speakers mostly work on French coming from simultaneous translation, as Parliament members use French only 1/3 of the time. Simultaneous translation is more difficult to repeat, due to hesitations and abrupt changes in speaking rate. Several trials have taken place since February 2002.

The third project, with the national sports network (RDS - *Réseau des sports*), has started in October 2005, and provides closed-captions for NHL Saturday night hockey games with Montréal’s Canadiens. Two topics are required, one for the description of game in action, and the other for more general interviews and reports between periods. Game action descriptions *must* be summarized by the shadow speakers, because captioning becomes unreadable at rates greater than 200 words per minute.

5.1. Shadow speaking results

We observed that a person with no previous experience attains a good performance in less than three weeks of practice with the system. Figure 4 was obtained when we introduced the new task of hockey game to our speakers; it shows a typical rate of progress for a new task or new shadow speakers.

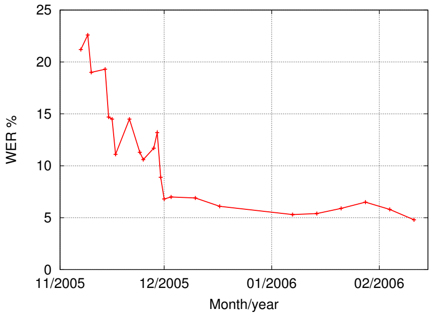


Figure 4: Word error rate progress for the RDS project.

Shadow speakers can work for relatively long periods without a break, depending on the task difficulty. In general they relay each other every 15 to 20 minutes.

Respoken speech is more difficult to recognize than read speech. We suspect phenomena like the Lombard effect, interference in speech planning, and wider speaking rate variations to be among the factors explaining this phenomenon. A 5% word error rate when reading typically jumps to 20% at first when respeaking.

In some applications, summarization is essential to making captions readable, but the additional cognitive load can reduce accuracy. Similarly, spontaneous speech in live reports or discussions may require a "simultaneous translation" to a language closer to the written form in order to be readable in captions.

Shadow speaking introduces a delay of 1 second or less.

5.2. Recognition accuracy

Table 1 summarizes the characteristics of current language models and vocabularies. In Table 2, we report recognition error rates following the usual convention of counting insertion, substitution and deletion errors relative to a reference text that corresponds to the words *spoken by the shadow speaker*. Errors due to homophony are counted, but capitalization errors are not. We did not explore the use of on-the-fly editing [8], although some other applications could tolerate a longer delay in exchange for error-free captions.

In all these projects, closed-captions were presented to panels of deaf and hard-of-hearing viewers and received good comments. As a rough comparison, real-time captioning of nine U.S. news programs was found to have an average word error rate of 11.9% using the same metric [9].

Topic	V_T	**OOV**	**G arcs**	**Ppx**
news_culture	38 K	1.3 %	889 K	114
news_economy	28 K	1.0 %	870 K	95
news_world	28 K	1.4 %	887 K	86
news_national	30 K	1.2 %	884 K	106
news_regional	30 K	1.3 %	884 K	106
news_sports	25 K	1.2 %	893 K	125
news_traffic	20 K	1.9 %	484 K	83
news_weather	20 K	1.4 %	448 K	72
hockey_inter	23 K	1.8 %	2.3 M	70
hockey_game	23 K	1.6 %	2.3 M	54
hoc_debates	21 K	1.0 %	2.4 M	55

Table 1: Vocabulary size, out-of-vocabulary rate, number of language model probabilities (G arcs) and test set perplexity.

Project	**Measurement period**	**Hours**	**Words**	**WER**
GTVA	11/2005 - 12/2005	26	83 K	11 %
HoC	09/2005 - 11/2005	20	163 K	8.8 %
RDS	12/2005 - 03/2006	33	195 K	7.2 %

Table 2: Results obtained during live trial periods.

6. Conclusion

Our speech recognition based captioning system has been deployed successfully in the course of several projects, and demonstrated its self-maintaining capability over long periods of time. Because shadow speakers can be trained in a short time, the system already is a viable solution for rapidly increasing the amount of closed-captioned programming. We are currently investigating its use to produce offline captions in quasi real-time.

7. Acknowledgments

This work was funded in part by GTVA, CANARIE Inc. and the Canada Heritage Fund for New Media Research Networks, and in partnership with the *Regroupement Québécois pour le sous-titrage*, an association which promotes deaf and hard of hearing rights to closed-captioning.

8. References

[1] J. Brousseau *et al*, "Automated closed-captioning of live TV broadcast news in French," in *Proc. Eurospeech*, Sept. 2–4, 2003, pp. 1245–1248, Geneva.

[2] R.E. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. (2/3), pp. 135–168, 2000.

[3] A. Matsui, H. Segi, A. Kobayashi, T. Imai, and A. Ando, "Speech recognition of broadcast sports news," Tech. Rep. No. 472, NHK Laboratories, 2001.

[4] A. Stolcke et al, "Entropy-based pruing of backoff language models," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 270–274.

[5] T. Niesler and D. Willet, "Unsupervised language model adaptation for lecture speech transcription," in *Proc. ICASSP*, May 2002, pp. 1413–1416, Orlando, Florida.

[6] 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. 171–185, 1995.

[7] J.-L. Gauvain and C. H. Lee, "Maximum *a posteriori* estimation for multivariate gaussian mixture observations of markov chains," in *IEEE Trans. SAP*, Apr. 1994, vol. 2, pp. 291–298.

[8] A. Ando, T. Imai, A. Kobayashi, H. Isono, and K. Nakabayashi, "Real-time transcription system for simultaneous subtitling of Japanese broadcast news programs," *IEEE Trans. Broadcasting*, vol. 46, no. 3, pp. 189–196, 2000.

[9] A. Martone, C. Taskiran, and E. Delp, "Automated closed-captioning using text alignment," in *Proc. SPIE Int. Conf. on Storage and Retrieval Methods and Applications for Multimedia*, 2004, pp. 108–116, San Jose, CA.