RECENT IMPROVEMENTS IN ESTONIAN LVCSR

*Tanel Alumäe**

Institute of Cybernetics at Tallinn Technical University
Estonia

ABSTRACT

This paper describes our current automatic transcription system for Estonian semi-spontaneous speech that we are developing within the Estonian language technology national program. A three pass decoding strategy is employed, with speaker-independent GMM acoustic models used in the first pass and speaker-adapted DNN-HMM models in the last pass. A neural network based phone duration model is used to rescore recognition lattices after the final pass and is found to give a surprisingly large gain in recognition accuracy. Compound words are split before building a statistical language model, and reconstructed from recognized hypotheses using an n-gram model. The word error rate of our system is 17.9% on broadcast conversations and 26.3% on conference speeches. This is around 8% absolute (24-30% relative) improvement compared to a GMM-based system of 2012.

Index Terms— Speech recognition, LVCSR, DNN, duration model, Estonian

1. INTRODUCTION

This paper describes our current offline speech-to-text transcription system for semi-spontaneous Estonian speech.

Estonian is the official language of Estonia, spoken natively by about one million people. It belongs to the Balto-Finnic branch of the Finno-Ugric language family. Speech recognition support for Estonian from commercial vendors practically doesn't exist. Even in fields that have pioneered in using speech recognition among other languages, such as medical dictation and automatic call routing, there aren't any commercial offerings for Estonian available. Google, known for its wide language technology coverage, has stated that it is not planning to implement support for Estonian in their speech recognition based products, such as voice search and voice typing[1].

The need for Estonian language technology has been recognized by the government and the field is now actively supported. In the context of the national program Estonian Language Technology 2011-2017 and its predecessor [1], over 100 hours of speech from various domains was manually transcribed with the purpose of improving speech recognition. Several large vocabulary continuous speech recognition (LVCSR) applications for Estonian were developed at the Institute of Cybernetics at Tallinn University of Technology. The Transcribed Speech Archive Browser [2] is a web application that provides access to hundreds of hours of automatically transcribed radio broadcasts (mainly conversational programs and long news broadcasts). End-users can use a web browser to navigate in the hierarchy of speech recordings, search from the transcriptions, view the transcriptions and listen to the recordings. Another application targeted towards end-users is the web-based speech transcription service[2]. The service allows users to transcribe long speech files by uploading them to the lab's server. The speech files are transcribed on the server and the resulting transcripts are sent back via e-mail. The service can also be used through a simple API. Another end-user product, *Diktofon*[3], is an application for the Android smart-phone platform that provides basic digital voice recorder functionalities. In addition, it uses the previously mentioned web-based speech transcription service to provide automatic transcripts for recordings containing Estonian speech. All the described applications are free for users.

This paper gives a detailed description of the multi-pass transcription system that serves the applications. Several improvements and redesigns have been made to the system since it was last described [3]. It is now based on the Kaldi toolkit [4] and uses deep neural network based hidden Markov models (DNN-HMMs) as main acoustic models. Neural network based phone duration models are used for rescoring the final lattices, resulting in significant improvement over the pure DNN-HMM system. The system is free and open-source[4]. In addition to being used as a backend transcription system for some experimental end-user applications, it is also daily used by three Estonian media monitoring companies for transcribing radio and TV broadcasts.

*This work was partly funded by the Estonian Ministry of Education and Research target-financed research theme no. 0140007s12 and through the project Estonian Speech Recognition System for Medical Applications.

[1] According to Pedro Moreno's keynote at SLTU 2012.

[2] http://bark.phon.ioc.ee/webtrans/

[3] http://play.google.com/store/apps/details?id=kaljurand_at_gmail_dot_com.diktofon

[4] http://github.com/alumae/kaldi-offline-transcriber

2. SYSTEM DESCRIPTION

2.1. Speech data

For training the acoustic models (AMs), we use the following wideband Estonian speech corpora, about 122 hours in total:

- the BABEL speech database containing about 9 hours of dictated speech;
- a corpus of Estonian broadcast news that contains mostly dictated speech, with some semi-spontaneous studio and telephone interviews (16 hours);
- a corpus of broadcast conversations consisting of various talk shows from three radio stations, mostly discussing political matters in a semi-spontaneous studio setting (19 hours);
- a corpus of semi-spontaneous interviews from radio news programs, discussing mainly daily news and current events; in most of the recordings, the interviewee talks over the telephone (33 hours);
- a corpus of local conference and lecture speeches, recorded using a close-talking microphone (37 hours);
- a corpus of studio-recorded spontaneous monologues and dialogues [5] (8 hours of speech)

As noted, part of the acoustic data contains telephone interviews from broadcast news programs. No special processing was applied for the telephone speech portion, i.e., the MFCC features were computed from the spectrum of the 16kHz signal. According to our experience, this does not cause problems – speech recognition performance on a test set containing such telephone interviews is comparable to the performance on studio interviews.

For tuning and measuring system performance, two different domains are used: conference speeches and broadcast conversations (BC), with separate development and test sets for both domains. The conference domain development and test sets both consist of three 20-minute presentation recordings from a local linguistics conference. The development set for the broadcast conversations domain contains four radio talk shows from 2009, each about 45 minutes, with 11 unique speakers in total. For testing, seven radio talk shows from 2011, each about 45 minutes (17 speakers in total) were used. None of the development and test data was used for training.

All broadcast speech and conference recording data used for training and testing are available from the authors of the paper for free for unrestricted use. The BABEL speech database is available from ELRA (ELRA-S0086) and the spontaneous speech corpus from the respective authors (restrictions may apply).

2.2. Acoustic models

The AM inventory contains 43 phoneme models, a silence/noise model and a garbage model that is used to absorb unintelligible and foreign language words during training. Although different noises and fillers are annotated in our training data at a relatively fine-grained level, they are all mapped to a single silence/noise model during training. We also merge palatalized and unpalatalized versions of several phonemes into single acoustic units, since it is difficult to derive the correct palatalization from the orthographic word forms. Estonian is a quantity language: all vowels and consonants, with some exceptions, can occur in short and long segmental duration. We create distinctive models for short and long variants of all phones except /j/. Estonian language has actually three distinctive quantity degrees: short, long and overlong [5]. However, the distinction between long and overlong duration is a property of the word foot rather than phone, and is thus difficult to model using purely segmental units. Therefore, we ignore this distinction in our acoustic models. The other reason behind such simplifications is the fact that the distinction between palatalized and unpalatalized phonemes as well as the distinction between the long and overlong quantity degree, is usually not needed for discrimination between orthographic word forms (i.e., palatalization and overlong quantity (with some exceptions) is not visible in orthography).

Acoustic model training is based on the Kaldi Switchboard training recipe. The triphone GMM models used in the first passes of decoding have 4000 decision tree leaves and 100 000 Gaussians. Speaker-adaptive training (SAT) with the MMI objective function is used. For the last decoding pass, we build a DNN-HMM hybrid system where the DNN is trained to provide posterior probability estimates for the HMM states [6]. Acoustic feature vectors for GMM models are obtained by splicing together 7 frames of 13-dimensional MFCCs and projecting them down to 40 dimensions using LDA. Speaker-based cepstral mean normalization is applied to the MFCCs. Speaker adaptive training is done by estimating a fMLLR transform for each speaker. For the DNNs, a nine frame context window of LDA-transformed MFCC features (4 frames at each side) is used at input, followed by a second LDA transform that retains 250 dimensions out of the initial 360. The DNNs have four hidden layers each with 1367 neurons, the output layer has 3166 units. DNNs are trained using cross-entropy cost and stochastic gradient descent.

2.3. Language model

Text data sources used for training the language models (LMs) are listed in Table 1. Most of the written language corpora are compiled at the University of Tartu [7]. In order to have up-to-date language data, we scrape additional web data from news portals and blogs. Finally, transcriptions of conversational broadcast data (talk shows, telephone in-

Table 1. Language model training data

Source	Documents	Tokens
Newspapers	655 847	204M
Web news portals	370 201	79M
Fiction, movie subtitles	140	38M
Magazines and journals	82 851	29M
Parliament transcripts	6024	15M
Social media (comments, blogs)	N/A	13M
Broadcast conversations	311	0.47M
Conference, lecture transcripts	54	0.26M
Broadcast news	219	0.13M
Total	1.1M	378M

terviews) and conference speeches are used as a sample of spoken language.

Before using the text data for LM training, text normalization is performed. Texts are tokenized, split into sentences and recapitalized, i.e., converted to a form where names and abbreviations are correctly capitalized while normal words at the beginning of sentences are written in lower case. Recapitalization serves two purposes: first, it improves the language model by allowing it to discriminate between proper names and nouns that have the same form but are used in very different contexts. Second, it enables us to easily produce correctly capitalized speech transcripts which is important for several applications where the system is currently used. For expanding numbers into words, a non-trivial approach is needed as the exact textual representation of each number depends on the inflection. However, the inflection of a number is usually not visible in orthography and is inferred from the context by human readers. To determine the inflection of a written number, we employ a semi-supervised machine learning approach similar to [8]: a support vector machine classifier is first built using the numbers that are already written as words in training texts, and the classifier is then used to determine the inflection for the rest of the numbers. Neighboring words and their suffixes are used as features for the classifier. This approach yields about 90% accuracy on a small development set randomly extracted from training data.

As Estonian is a heavily compounding and inflective language, the lexical variety of the language is very high. To reduce the out-of-vocabulary (OOV) rate of the LM, compound words are decomposed into compound segments, using the word structure information assigned by the morphological analyzer [9]. The relationship between the vocabulary size, vocabulary type and OOV rate is shown on Figure 1. While it is clear that decomposing compounds helps to decrease OOV rate, it is perhaps surprising that there is such a big difference in the full word OOV rate between the BC and conference domains. This could be explained by the relative shortage of suitable LM training data for the conference domain, compared to the BC domain.

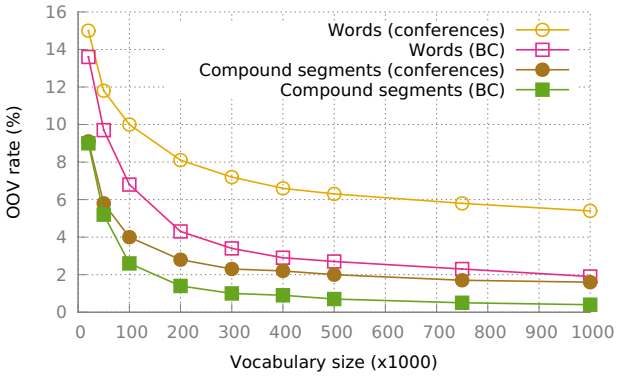

Fig. 1. OOV rates on the broadcast conversation (BC) and conference speech development sets, using full words and decomposed compounds as vocabulary items. Vocabularies were optimized for the particular domain using interpolation of domain-specific unigram counts.

Table 2. Out-of-vocabulary rates and language model perplexities, using a 200K vocabulary of with decomposed compounds.

	Broadcast conv.		Conference speeches	
	Dev	Test	Dev	Test
OOV	1.4%	1.0%	2.8%	3.1%
Perplexity	480	388	476	526

LM vocabulary is created by selecting the 200 000 most likely case-sensitive compound-split units from the unigram mixture of the corpora, optimized on the development texts. For each corpus, a 4-gram LM is built. The LMs are compiled by including all bigrams and trigrams as well as 4-grams occurring more than once, using interpolated modified Kneser-Ney discounting. The individual LMs are interpolated into one by using interpolation weights optimized on development data. Finally, the LM is heavily pruned to less than one tenth in size using entropy pruning. This is all done using the SRILM toolkit [10]. Since the vocabulary of LM is relatively large, an even more aggressively pruned LM is created for the finite state transducer (FST) cascade used at decoding.

For the current work, two different LMs are built: for conference speeches and for conversational broadcast speech. For both domains, the LM is optimized on the development data of the corresponding domain (see section 2.1). The out-of-vocabulary (OOV) and perplexity results of the pruned LMs are listed in Table 2.

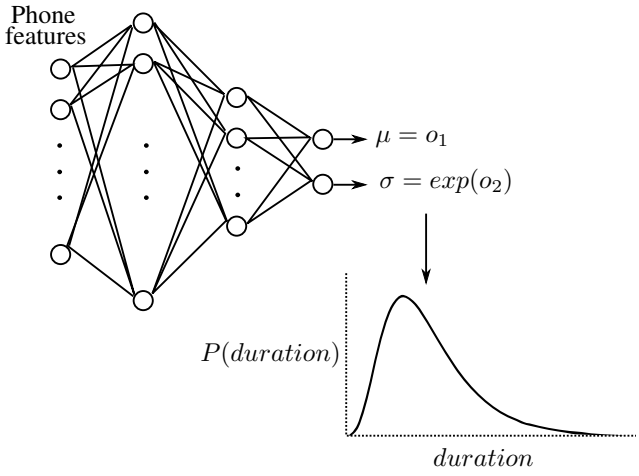

Fig. 2. Architecture of the neural network phone duration model. Input features describe the current phoneme, phonetic context of the current phone, and the durations of the previous phones. Output is used for calculating the log-normal duration probability density function of the current phone.

2.4. Pronunciation dictionary

As Estonian is a language with a close relationship between word orthography and pronunciation, a rule based system is used for deriving the pronunciations for words in the LM lexicon [5]. The rules are mainly concerned with determining the correct variant (short or (over)long) of the plosives based on the usage context. For many common foreign proper names and abbreviations, pronunciation is created by first transforming the lexical form to a localized form using a transliteration table, and then applying the common pronunciation rules.

2.5. Phone duration model

As noted in section 2.2, Estonian phonemic inventory shows a contrastive opposition of short and long vowels and most of the consonants. HMMs model phone durations using state transition probabilities, resulting in geometric probability density functions (PDFs) for phone durations. However, phone durations are known to follow a gamma or log-normal distribution, rather than geometric. Several methods have been proposed to resolve this conflict. Improved duration modeling can be integrated directly into the HMM framework by replacing the HMM state transition probabilities with explicit duration PDFs [11] or by modifying the HMM topology [12]. An alternative approach is to model word or phone durations using an independent model and use it as a separate knowledge source during N-best rescoring [13] or lattice rescoring [14, 15]. Such approaches usually result in a small (usually about 5% relative) word error rate reduction in large vocabulary continuous speech recognition (LVCSR) tasks.

Recently, we proposed a new method for modeling phone durations in speech recognition [16]. The model was based on a decision tree that finds clusters of phones in various contexts that have similar durations. For each resulting phone cluster, a log-normal duration probability density function (PDF) was estimated that was used for N-best rescoring.

We have now improved the previous model by replacing the decision tree based binned probability model with a neural network that computes the parameters (mean and standard deviation) of the log-normal phone duration PDF from the phone's contextual features.

More specifically, the duration model is a neural network with two hidden layers (Figure 2). Inputs to the network consist of mostly binary features of the phone and its neighboring phones, such as the phoneme label and type, position in the word and sentence (e.g., 'is the phone word-final?'), whether the phone is pre-pausal or post-pausal, phonemic length (short or long), position of the current syllable in the word. To better model varying and non-stationary speaking rates the contextual features also include the observed duration values of previous phones, normalized non-linearly to the $(0, 1)$ range. The network has two outputs: one for the conditional mean of the duration PDF and second for the conditional standard deviation. Because the predictions of the network correspond to the parameters of the log-normal distribution, exponent function is applied to the output corresponding to the standard deviation to guarantee that its positivity.

In the reported experiments we use a context window of size 7 (3 phones from either side) which results in 406 input features to the neural network. The first hidden layer has 600 units and uses rectified linear activation functions. The second hidden layer has 300 units and applies the *maxout* activation function [17]. We regularize the model by imposing a constraint on the norm of each hidden layer weight vectors.

The model is trained using back-propagation using negative log-likelihood of the data as the cost function. Features and duration values obtained from a large phonetically aligned speech corpus are used as training data. Once trained, the outputs of the model (mean and standard deviation) are used to estimate the likelihoods of phone durations in the recognition lattices. The resulting phone duration likelihoods can then be applied for lattice rescoring.

A more detailed description of the neural network based phone duration model, including experiments with other languages, is already submitted for publishing.

2.6. Decoding strategy

Transcribing a speech recording in our system consists of the following steps: speech detection and speaker diarization, followed by three passes of decoding with increasingly accurate acoustic models, language model rescoring, phone duration

[5] Available at `http://github.com/alumae/et-g2p`

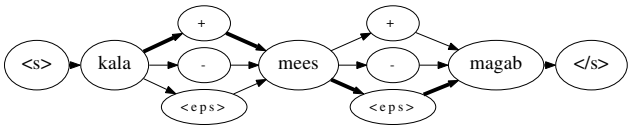

Fig. 3. A finite state automaton representing all possible compound word reconstructions of the token sequence "kala mees magab" (Eng. lit. "fisher man sleeps"). Correct path, corresponding to the words "kalamees magab" ("fisherman sleeps"), is drawn in bold.

model rescoring and compound word reconstruction.

Input audio data is segmented into shorter sentence like chunks using the LIUM SpkDiarization [18] toolkit. Segments are classified as speech or non-speech using a Gaussian mixture model built from our AM training data. Segments containing speech are clustered, each cluster corresponding ideally to one unique speaker in the recording. BIC clustering followed by CLR-like clustering [19] are applied. The resulting speaker labels are used to perform unsupervised AM adaptation in the multi-pass decoding system.

In the first decoding pass, speaker-independent GMM-HMM AMs are used. The resulting hypotheses are used for creating fMLLR adaptation transforms for each speaker which are then used in the second pass with speaker-adaptively trained GMMs. The results are used to re-estimate the speaker fMLLR transforms. In the final decoding pass, speaker-adaptive DNN-HMMs are used to generate lattices for each utterance. The lattices are rescored using a larger language model.

From the resulting lattices we generate new lattices that contain phone segmentation information for each word edge. We then use the phone duration model to add log duration scores for each word edge in the lattice, by summing the log duration likelihoods of the word's phones. An additional constant equal to the number of phones in the word is added to each word edge to be used as a phone insertion penalty. For edges representing silence or noise, the duration and phone penalty scores are set to zero. Finally, duration model score, phone penalty, LM score and AM score are combined, using weights optimized on the corresponding development sets, and the lattices are decoded.

2.7. Compound word reconstruction

Output hypotheses of the decoding system consist of sequences of word-like units where compound words (including words containing hyphen as a separator) are replaced with their compound segments. In the final system output, compound words have to be reconstructed from the segments. To reconstruct compound words, we use a weighted finite state automaton (FSA) based approach. The input tokens are used to compile a FSA with three possible paths between each consequent token (see Figure 3). The paths between the tokens correspond to a compound word ("+" on the figure, meaning that the neighboring tokens should be concatenated), a dash-connected word ("-", as in võib-olla, Eng lit may-be), and a non-compound ("<eps>", corresponding to an epsilon node, meaning that there is a space between the tokens). The FSA is then composed with a 4-gram language model, represented as another FSA, that is trained on texts with the respective special tokens ("+" and "-") between compound word and dash-compounded word segments. Shortest path in the resulting FSA gives the most likely reconstruction of the compound words.

Table 3. Word error rates in percentages of the transcription system for the broadcast conversations and conference presentation domains. Accuracy of the GMM-HMM system developed in 2012 and described in [3] is given for comparison.

	BC		Conferences	
	Dev	Test	Dev	Test
GMM-HMM ML (2012)	24.9	25.6	31.5	34.6
GMM-HMM SAT+MMI	22.6	22.7	27.9	30.0
DNN-HMM	20.7	20.0	24.9	27.9
DNN-HMM + dur model	18.0	17.9	23.7	26.3

3. EXPERIMENTAL RESULTS

Word error rates (WER) of the system on two test domains are reported in Table 3. The first line lists the WER results of our previous system built in 2012 and described in [3]. The system uses maximum-likelihood trained GMM-HMMs, and performs both unsupervised fMLLR and MLLR adaptation as well as confusion decoding. It also uses slightly less data for training acoustic and language models. The second line gives the WER of the current system after decoding with GMM-HMM acoustic models. Unlike the 2012 system, they are trained using MMI and SAT, and two passes of fMLLR adaptation are performed. The third line corresponds to the speaker-adaptive DNN-HMM acoustic models. The last line lists the results after rescoring lattices with the phone duration model.

The improvements of the current GMM-HMM and DNN-HMM system over the previous system are not surprising: both MMI training and DNN-HMM models are expected to give roughly 10% relative improvement in WER. Somewhat unexpected is the improvement brought by the phone duration model: it gives 6-11% relative reduction in WER over the DNN-HMM system that is already highly optimized. Overall, the reduction in WER over the 2012 system is 7.8% (30% relative) and 8.3% (24% relative) for the two test sets.

4. CONCLUSION

The paper described our Estonian speech-to-text system developed with the aim of providing accurate offline speech transcription service and software to end-users and various institutions.

The system uses the Kaldi toolkit and is thus able to use various state-of-the art acoustic modeling techniques, such as DNN-HMM models. The resulting improvements over our previous GMM-based system are consistent with previous work reported in the literature. On top of the DNN-HMM system, we apply a novel neural network based phone duration model that gives further 1.6-2.1% absolute WER reduction. The final WER of the system is 17.9% for broadcast conversations and 26.3% for oral conference presentations. The system is free, open source, and is used on a daily basis by several local media monitoring companies.

Future work on this system includes using more advanced feature representations and training criteria for the DNN-HMM models and using improved language modeling techniques, such as recurrent neural network LMs.

5. REFERENCES

[1] Einar Meister, Jaak Vilo, and Neeme Kahusk, "National programme for Estonian language technology: a pre-final summary," in *Baltic HLT 2010*, 2010, pp. 11–14.

[2] Tanel Alumäe and Ahti Kitsik, "TSAB – web interface for transcribed speech collections," in *Interspeech 2011*, Florence, Italy, 2011, pp. 3335–3336.

[3] Tanel Alumäe, "Transcription system for semi-spontaneous Estonian speech," in *Baltic HLT 2012*, Tartu, Estonia, 2012.

[4] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The Kaldi speech recognition toolkit," in *IEEE ASRU Workshop*, Dec. 2011.

[5] Pärtel Lippus, *The acoustic features and perception of the Estonian quantity system*, Ph.D. thesis, Tartu University, 2011.

[6] Frank Seide, Gang Li, and Dong Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Interspeech 2011*, Florence, Italy, 2011, pp. 437–440.

[7] Heiki-Jaan Kaalep and Kadri Muischnek, "The corpora of Estonian at the University of Tartu: the current situation," in *Baltic HLT 2005*, Tallinn, Estonia, 2005, pp. 267–272.

[8] Richard Sproat, "Lightly supervised learning of text normalization: Russian number names," in *SLT 2010*, Berkeley, California, USA, 2010, pp. 436–441.

[9] Heiki-Jaan Kaalep and Tarmo Vaino, "Complete morphological analysis in the linguist's toolbox," in *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, Tartu, Estonia, 2001, pp. 9–16.

[10] Andreas Stolcke, "SRILM – an extensible language modeling toolkit," in *Proceedings of ICSLP*, Denver, USA, 2002, vol. 2.

[11] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Language*, vol. 1, no. 1, pp. 29–45, 1986.

[12] M. J. Russell and R. K. Moore, "Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition," in *ICASSP 1985*, 1985.

[13] A. Anastasakos, R. Schwartz, and Han Shu, "Duration modeling in large vocabulary speech recognition," in *ICASSP 1995*, 1995, vol. 1, pp. 628–631.

[14] Dino Seppi, Daniele Falavigna, Georg Stemmer, and Roberto Gretter, "Word duration modeling for word graph rescoring in LVCSR," in *Interspeech 2007*, 2007, pp. 1805–1808.

[15] N. Jennequin and J.-L. Gauvain, "Modeling duration via lattice rescoring," in *ICASSP 2007*, Honolulu, HI, USA, 2007, vol. 4, pp. 641–644.

[16] Tanel Alumäe and Rena Nemoto, "Phone duration modeling using clustering of rich contexts," in *Interspeech 2013*, Lyon, France, 2013.

[17] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," *ArXiv e-prints*, Feb. 2013.

[18] S. Meignier and T. Merlin, "LIUM SpkDiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, Dallas, TX, USA, 2010.

[19] C. Barras, Xuan Zhu, S. Meignier, and J. L. Gauvain, "Multistage speaker diarization of broadcast news," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1505–1512, Aug. 2006.