# Automatic Word Decompounding for ASR in a Morphologically Rich Language: Application to Amharic

Thomas Pellegrini and Lori Lamel, *Member, IEEE*

*Abstract*—**This paper investigates a data-driven word decompounding algorithm for use in automatic speech recognition. An existing algorithm, called "Morfessor," has been enhanced in order to address the problem of increased phonetic confusability arising from word decompounding by incorporating phonetic properties and some constraints on recognition units derived from forced alignments experiments. Speech recognition experiments have been carried out on a broadcast news task for the Amharic language to validate the approach. The out of vocabulary (OOV) word rates were reduced by 35% to 50% and a small reduction in word error rate (WER) has been achieved. The algorithm is relatively language independent and requires minimal adaptation to be applied to other languages.**

*Index Terms*—**Automatic speech recognition (ASR), broadcast news transcription, less-represented languages, lexical modeling, morphologically rich languages (MRLs).**

## I. Introduction

In the literature, languages such as Arabic, Finnish, Turkish, and Estonian, are often referred to as "morphologically rich languages" (MRLs). Other languages do not have a "poor" morphology, this qualification emphasizes the highly productive processes involved in word formation in MRLs. For such languages, it is common to generate words by the compounding of smaller units that are primarily lexical morphemes (such as in German), or mostly grammatical morphemes (for example, Semitic languages such as Arabic or Amharic), or both (such as Turkish). These languages need very large lexicons, containing several hundred thousand words, to achieve good lexical coverage. Since state-of-the-art automatic speech recognition (ASR) systems generally use fixed (also called closed) lexicons, only the words in the recognition lexicon can potentially be recognized. For MRLs, the rich morphology implies a high number of unknown or out-of-vocabulary (OOV) words, which typically produce 1.5 to 2 errors for each OOV word [1]. This large lexical variety also poses a problem for language modeling, where it can be difficult to have reliable n-gram estimates for infrequent words. To address these issues, word decomposition has been investigated in a number of studies for various languages such as German [2], [3], Turkish, Finnish and Estonian [4], Vietnamese [5], and Dutch [6]. The probabilistic word decomposition framework used in this study is derived from the baseline version of the corpus-based word decompounding algorithm "Morfessor" [7].

High OOV rates and poor language model estimation are problems that also arise when developing technologies for less-represented languages, for which little data are available in an electronic form. Most of the world's languages suffer from poor representation on the web, which is being used more and more as the primary source for collecting data (principally texts) for building ASR systems [8]. This study reports on experiments carried out with the Amharic language, the official language of Ethiopia, which is both a less-represented language and a language in which grammatical compounding is frequent [9]. For morphologically rich languages and less-resourced languages, the first issue to be addressed, is the high percentage of unseen words as typical OOV rates are higher than 7%. Previous work reported improvements in ASR for Amharic broadcast news data when using sub-word units: for a relative OOV reduction of 16%, a 10% relative reduction in word error rate (WER) was achieved [10]. The sub-word units were identified with a character-based maximum branching factor algorithm similar to the one used in [2], and selected using a heuristic. In the same study, it was shown that experiments allowing more decompositions led to increased insertion and deletion rates, and to an overall degradation in performance (a 7% relative increase in WER, with a 20% relative OOV reduction compared to the best sub-word based system). A common observation in the literature is that small lexical units can often be less reliably decoded than longer units, since these units are acoustically more similar and therefore more confusion-prone [11]. One solution to overcome the increased confusion, consists of using word-based models to generate N-best lists or lattices, and a sub-word unit language model, only in a final rescoring framework. Nevertheless, for several reasons it was chosen to investigate the use of sub-word units in all stages of the decoder. First, Amharic is a language that has a rather straightforward grapheme-to-phoneme conversion, allowing pronunciations to be easily produced for sub-word units [12]. Second, the use of the same lexical units in all steps of the decoding simplifies the global process. Finally, we wanted to investigate new features that try to incorporate "oral" properties in the identification/selection of the sub-word units, in an attempt to take account of some specificities of spoken language. One of these new properties is based

Manuscript received August 04, 2008; revised April 10, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tanja Schultz.

The authors are with LIMSI-CNRS, BP133, 91403 Orsay cedex, France (e-mail: thomas@l2f-inesc.pt, lamel@limsi.fr, thomas@l2f-inesc.pt, lamel@limsi.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2022295

on the distinctive features specific to the Amharic phonemes. By giving a "phonemic" distance between two lexical units, word splits that result in the largest distances between sub-word units can be favored. The distinctive features are based on very general theoretical sound properties. According to Jakobson, phonemes of a specific language are distinguishable with a small set of articulatory and acoustico–perceptive features, called *distinctive features*, such as voiced-unvoiced property or the place of articulation often corresponding to the point of constriction in the vocal tract [13]. A problem is that splitting words can create homophones or near-homophones, particularly if multiple pronunciation variants are allowed for the lexical units. To overcome this drawback, an additional constraint was introduced to forbid word splits that could have the same pronunciation variant. Results incorporating these properties for vowels were reported in [14], showing an absolute WER reduction of 0.4% relative to the word-based system. However, the experiments in this previous work were carried out on a development corpus, since no additional test corpus was available. In the present article, cross-validation has been used to test the approach. The distinctive feature property has also been extended to the consonants, while in previous work it was limited to the vowels. Finally, complementary experiments with longer-span language models (5-gram) are also reported.

The paper is organized as follows. The next section discusses the key role of words in ASR and motivates the use of sub-word units for MRLs. This is followed by an overview of the ASR literature with sub-word units. Section IV describes the baseline version of the corpus-based word decompounding algorithm Morfessor, with the modifications made to incorporate "oral properties." Section V presents the experimental results carried out on the Amharic corpus. Since morphological decomposition results in the redefinition of words or lexical entries used for ASR, each explored configuration implies renormalization of the available texts and transcripts, as well as the retraining of the language and acoustic models. All modifications in the word decompounding algorithm are fully tested by measuring ASR performance in terms of word error rates in comparison to the reference word-based system. Finally, some conclusions and perspectives are given.

## II. Reference Units for Speech Recognition

Speech recognition consists of finding the best elementary unit sequence $\hat{M}$, which is the hypothesis with the highest probability, given a speech signal $S$: $\hat{M} = \arg\max_M P(M|S) = \arg\max_M P(S|M)P(M)$. By and large, the most widely used recognition unit is the "word," where the definition of a word may vary across languages and systems. Performance is usually measured by word error rate, which is the sum of all kinds of word errors (insertions ($\#I$), substitutions ($\#S$), and deletions ($\#D$)), normalized by the number of words ($N$) in the reference (manual transcription in general). The WER is formulated as $\mathrm{WER} = (\#I + \#S + \#D)/N$. Word errors are determined by dynamically aligning the recognition hypothesis to the reference transcription at a sentence level. We used the NIST standard scoring tool "sclite," available at http://www.nist.gov/speech/tools.

In Linguistics, the concept of *word* is often described as complex and problematic, with difficulties arising when word identification has to be done.[1] In speech recognition, only words specified in a lexicon can be recognized. So some kind of word segmentation and identification are necessary to build a recognition lexicon, and it is typical to take a very pragmatic approach, identifying words in as simple a manner as possible. Even for languages written with a space or another separator between words, there are normalization choices to make. In French for example, the use of the apostrophe is very frequent, as for the definite article $l'$. Words like $l'oral$ can be considered as two words $l'$ and *oral*, or just one word since there is no space between the two distinct words. In order to avoid increasing substantially the lexicon size, the first possibility may be chosen, and all small words $(c', j', l', m', n', s't', \ldots)$ may be separated from their associated nouns and considered as words. This choice reduces lexical variation at the cost of introducing many words with a single phone. Such normalization issues in French are discussed in [16]. As explained in the Chapter "The use of lexica in Automatic Speech Recognition," by Adda-Decker and Lamel [17], normalization choices for the apostrophe may be different in French and in English. In English, apostrophes are not as frequent as they are in French, and therefore they are typically not considered to be word separators. Contractions like *I'll*, *you've*, or *he's* and as well as compound words and multi-word sequences are often used as lexical entries for speech recognition [18]–[20]. These normalization practices, derived from experience gained by specialists in speech recognition, may be different according to the experts that choose them, but they illustrate well the issues linked with word definition for ASR. The specific choices may also differ depending on the language, task, and application. Dialog systems and conversational speech recognizers have been reported to benefit from using compound words in order to facilitate the use of pronunciation variants specific to conversational speech.

Some languages have no word separators, as it is the case for various Asiatic languages such Chinese, Japanese, and Thai. For these languages, segmentation algorithms are required for pre-processing and/or postprocessing. In general, a reference lexicon is used but very often, multiple word segmentations are possible for the same sentence. Various automatic techniques have been proposed to try to remove this ambiguity, the most popular being the maximum match segmentation, which tries to find the longest words to match the characters in a sentence. In 2005, the "Second International Chinese Word Segmentation Bakeoff" showed that despite performance gains in the word segmentation task, the main issue is still the processing of the OOV words [21]. In order to avoid this issue, the ASR performance is typically measured at character level, with character error rates (CERs) instead of word error rates [22], [23].

For morphologically rich languages, the definition and selection of lexical units is a popular topic in ASR, since prohibitive lexicon sizes would be required to achieve reasonable

[1] Linguists use other concepts, such as *word-form*, *lexeme*, and *autonomous syntagm* [15], for example.

TABLE I
OUT-OF-VOCABULARY RATE (OOV) COMPARISON FOR TWO RICH MORPHOLOGY LANGUAGES (AMHARIC AND TURKISH), AND TWO LANGUAGES THAT HAVE A "LESS RICH" MORPHOLOGY (ENGLISH AND FRENCH)

| *Language* | *Lexicon size (word types)* | *OOV(%)* |
|---|---|---|
| *English* | 65k | 0.6 |
| *French* | 65k | 1.2 |
| *Amharic* | 133k | 6.9 |
| *Turkish* | 250k | 6.5 |

lexical coverage. One characteristic of such languages is the increase in distinct word number ("word types"), as a function of the total number of words of a corpus ("word tokens"). For these languages, the increase is much faster than for other languages. The study reported in [24] for example, distinguishes Finnish, Estonian, Turkish, and Arabic from English on that point. Table I gives lexicon sizes and OOV rates of systems developed at LIMSI for English and French, and for two morphologically rich languages, Amharic and Turkish. Nowadays, it is common practice to use lexicons comprised of at least 65k words and most state-of-the-art recognition system developers consider acceptable OOV rates to be under 1%. As shown in Table I with 65k words the OOV rate for English is 0.6%, and is on the order of 1.2% for French. Using a 200k word lexicon can reduce the OOV rate to under 0.5% for French [25].

For Amharic and Turkish, much higher OOV rates are observed, 6.5% and 6.9% respectively, with substantially larger lexicons. This difference is mainly due to the rich morphology of Amharic and Turkish, but is also accentuated by the lack of resources compared to English and French. In [26], a 96.4 million word text corpus is used to train language models for broadcast news transcription in Turkish. If all observed word forms were included in the lexicon, it would be comprised of 1.4M words, a prohibitive size for speech recognition. Lexicon size reduction is quite interesting in that case, and decompounding words into sub-word units can serve to decrease both the recognition lexicon size and OOV rates. Some illustrations found in the literature are as follows.

- The German word *Schulelternbeiratsmitglieder* was decompounded into *Schuleltern + beiratsmitglieder*, then into *Schul + eltern + beirats + mitglieder*, by using a character-based maximum branching factor algorithm [2].
- The Turkish sentence *Isteklerimizi elde ettik dedi* has been decompounded into *Istekler+ imizi el+ de etti+ k de+ di*, by using the Morfessor algorithm, also used in this work and presented in Section IV[27].

Recently, several sites have reported on morphological decomposition for the Arabic language [28]–[30] where sub-word units such as prefixes (*Al, bAl, fAl, kAl, b, f, k, l, s, w, . . .*) are used to decompound words. Rules are typically applied to restrict the decomposition of frequent words avoiding some possible confusions. These reported experiments were carried out with state-of-the-art systems trained on very large corpora.

Some of the above-mentioned studies showed improvement in recognition performance obtained by word decompounding. The methods used in these studies were different, Section III presents and discusses some of them, along with other studies found in the literature. Here first is a brief introduction to the Amharic language, which is used as a case study in this work.

TABLE II
EXAMPLE OF DIFFERENT ORDERS (SYLLABLE NUCLEUS) ASSOCIATED TO THE 'L' CONSONANT, GIVEN IN AMHARIC SCRIPT AND OUR LATIN TRANSLITERATION. THE '$x$' STANDS FOR A REDUCED VOWEL (SCHWA)

| Ge'ez Symbols | ለ | ሉ | ሊ | ላ | ሌ | ል | ሎ |
|---|---|---|---|---|---|---|---|
| Transliteration | lE | lu | li | la | le | lx | lo |

Amharic was chosen as an example of a Semitic language, language family to which Arabic belongs to. It is mainly spoken in Ethiopia. After Arabic, it is the second most widely spoken Semitic language in the world, with 22 million speakers [31]. Despite its "official working" language status, and its nationwide use, Amharic suffers from poor representation on the Internet, and may be considered as a "less-represented" language, for which only small quantities of written texts are available [32]. For speech recognition, the lack of text resources makes language model probability estimation difficult, and often implies high out-of-vocabulary rates. In Amharic, these problems are increased by its rich and complex morphology, which is inflectional and derivational [33]. One characteristic of languages with a rich morphology is a high increase in the number of word types as a function of the number of word tokens [34]. Reference [9, Table IV] compares the frequencies of word types in Amharic and in English, showing that word type frequencies are quite a bit lower for Amharic.

Amharic has 34 basic symbols, for which there are seven vocalizations (transliterated form): /ɛ/, /u/, /i/, /a/, /e/, /ə/ and /o/,, referred to as the seven orders. The basic symbols are modified in a number of different ways to indicate the different vocalizations. 85% of the syllables represent a CV sequence (C for consonant and V for vowel), one symbol represents the complex sound /ts/V and the remainder represent CwV sequences (where w is a semi-consonant). In this study, Cw has been considered as a single phone. For practical reasons, the Amharic script was transliterated into a set of Latin letters. Table II shows an example of the ለ syllable, that is transliterated by /lE/ corresponding to the phonetic transcription [lϵ], given with its seven orders. The sixth-order syllable nucleus is a schwa, written as "x."

## III. Word Decompounding for Speech Recognition

The use of sub-word units in speech recognition is not new, with studies dating from the mid 1990s, but it remains an active research area. Most of the studies use "Top-Down" methods: starting from full word forms, words are decompounded into smaller units. Once sub-word units have been selected, the studies differ on how the sub-word units are used in the decoding. Sub-word units can be used at different levels of modeling: acoustic modeling and/or language modeling, for all the decoding, or just during lattice rescoring. Kirchhoff and Sarikaya, who led the ISCA Workshop "Processing Morphologically Rich Languages" during the Interspeech conference in Antwerp, August 2007, distinguished three main approaches that are briefly reviewed here [34].

To better identify the pros and cons of the three methods, they are presented in relation to the generic speech recognition formulation as shown in (1). The likelihood $P(S|M)$ of the signal $S$ given the word sequence $M$, is developed as the sum of the pronunciation probabilities (pronunciation variants named $H$) associated to the word sequence $M$. $H$ corresponds to a sequence of acoustic models (phones in general)

$$\begin{aligned}\hat{M} &= \arg\max_{M} P(S|M)P(M) \\ &= \arg\max_{M} \sum_{H} P(S|H)P(H|M)P(M). \qquad (1)\end{aligned}$$

The first method consists of using sub-word units in all modeling elements of the speech recognizer: acoustic models, lexicon, and language model. This is the approach taken in [3], [35], for example, with application to German and French. If the French word *aller* was decompounded into *all- er*, then with this method, two acoustic models would be used, one for *all-* and one for *er*. The advantage here is the economy of the pronunciation lexicon, since in this case $P(H|M) = P(M|M) = 1$. A limitation of this method is the high complexity of the acoustic models, with a number of states that should depend on the number of phones in each sub-unit.

The second method uses the sub-word units in the lexicon and the language model in the decoding process. The recognition units may be a combination of words and sub-word units. Acoustic units are not based on the sub-word units, but are generally phones or phone-like units. Some studies using this approach are for example [4], [24], [28], [29], [36]. With this method, building a pronunciation lexicon (i.e., determining $H$ for each $M$) is necessary and may pose difficulties for some sub-word units. One solution would be to decompound words into sub-word units for which pronunciations are known or easy to determine. In the literature, languages for which this method has been used have a simple grapheme-to-phoneme conversion. This is globally the case for the Amharic language of interest in this study. Another potential problem source is modeling the $P(S|H)$ term, in the sense that small units are known to increase acoustic–phonetic confusions for the system, and their probabilities $P(S|H)$ are very similar. In the present work, special care in sub-unit generation was taken in order to try to avoid the creation of units that were too small or too similar, with the use of the new "oral" properties, presented in Section IV-C.

The third method uses sub-word units only in a rescoring pass, i.e., a sub-word based language model is used to rescore recognition hypotheses, generally structured as lattices or consensus networks, generated by a word-based system. Examples of this method can be found in [11], [27], [37]. Again concerning (1), only $P(M)$ is modified in comparison to a word-based system. The advantage of this method is double: there is no increase in acoustic–phonetic confusability since words are used during the acoustic part of the decoding, and there is no problem of finding pronunciations for sub-word units. Various strategies can be adopted: simply decompounding words from the N-best hypotheses and rescoring with a sub-unit based language model (LM); combining scores achieved with a word based LM and scores achieved with a sub-unit based LM [11]; or expanding lattices or consensus networks by adding nodes and arcs with words that begin with a same prefix for example [27].

Based on the literature studies, it is not possible to determine which of the three approaches is the best, and the end choice is likely to depend on a variety of factors and constraints. Since the work presented here is for a language that has a straightforward grapheme to phoneme conversion, the second approach, that combines the use of sub-word units for language modeling in all decoding steps with phone-based acoustic units, was chosen. In Section IV, the word decompounding strategy, enhanced for speech recognition purposes, is described.

## IV. Incorporating ASR-Oriented Properties in Corpus-Based Word Decompounding

Automatic word decompounding is investigated as a means to help select recognition units in an almost language-independent manner. In order to minimize the work needed to apply the adopted approach to different languages, a data-driven algorithm, requiring little linguistic knowledge, was explored. Various unsupervised morphology analysis algorithms are open source or easy to implement, such as Harris [38], Goldsmith [39], and Morfessor [7]. The Morfessor algorithm was chosen since it seemed to be a more general model than the others, for example unlike Goldsmith, no assumption about the basic structure of words is made. Furthermore, several recent studies making use of Morfessor reported improvements for a variety of languages, using either the second [4], [24], or the third [27] approach described in the previous section. This work is an extension of the Morfessor algorithm, as implemented in the open source Perl program called "Morfessor 1.0," available at http://www.cis.hut.fi/projects/morpho/.

### A. Baseline Morfessor 1.0 Algorithm

Morfessor is an iterative algorithm that given a corpus, proposes word segmentations found with an optimization criterion. The authors use the term of "morphemes" to name the sub-word units proposed by Morfessor, but they also use the neologism "morphs," since the splits are not always true morphemes in a linguistic sense. Finally, morphs can be either words or word splits.

An overview of the basics of this algorithm is provided here, for further information the reader is referred to [7]. The program has two modes:

1) A "training" mode which creates a word segmentation model given a lexicon with optional frequency counts. Training uses a maximum a posteriori (MAP) criterion based on several text properties, including word frequencies and string probabilities.
2) A "decoding" mode in which a previously learnt decomposition model can be used to decompound a new word list. Each input word is decomposed into a sequence of morphs that exist in the model. This search algorithm maximizes only the morph frequencies, and no retraining is done. Words that are not in the model can be decomposed into a sequence of known morphs.

During model training, the algorithm tries to iteratively maximize the following estimate:

$$M = \arg\max_{L} P(L|\text{corpus}) = \arg\max_{L} \underbrace{P(\text{corpus}|L)}_{\text{Likelihood}} \underbrace{P(L)}_{A\ priori} \quad (2)$$

where $P(\text{corpus}|L)$ is the maximum-likelihood estimate of the corpus given a lexicon $L$, based on the word frequencies, and $P(L)$ is the *a priori* probability of the lexicon $L$, i.e., the probability of getting $M$ distinct morphs $m_1, \ldots, m_M$ as shown in (3). Properties used in the baseline version are morph frequency, morph string, and morph length, respectively, denominated $n(m_k)$, $s(m_1)$, and $l(m_1)$ in (3). For more details about the computation of these terms, the reader is referred to [7]. Our modifications, described in the following sections, affect the *a priori* properties used as

$$P(L) = P(n(m_1), \ldots, n(m_M))P(s(m_1), \ldots, s(m_M))P(l(m_1), \ldots, l(m_M)). \quad (3)$$

As it is common practice for this type of algorithm, probabilities are not multiplied as is, since they are often very small, but the negative log probabilities are summed. Maximizing the likelihood consists then in minimizing a sum of negative log probabilities, which can be seen as minimizing a cost function.

The decoding part of Morfessor is different from the training mode, since chosen morphs are those which minimize a cost function based only on the morph frequencies, and no other property.

In both modes, every word position is a potential candidate for split, and the algorithm explores all word substrings. Words can be split into various morphs, but words are not decompounded if splitting does not reduce the cost function value.

### B. Modified End-of-Word Probability

In the baseline Morfessor program, the character probabilities are static constants, calculated only once during model initialization, as the simple ratio of the number of occurrences of the character divided by the total number of characters in the corpus. These are independent of word position. To represent the word boundary, a space character is added to each lexical entry. The end-of-word probability is the probability of the space character, and has the same value for all words and morphs in the corpus.

Inspired by Harris' algorithm [38] and previous work on German word decomposition [2], we propose replacing this static probability by the probability $P_H$ defined in (4), to take the string context into account. $P(l(m_1), \ldots, l(m_M))$ in (3) is replaced by $P_H(l(m_1), \ldots, l(m_M))$. The word beginning symbol (WB) stands for the strings that begin a given word, from length zero to the word length itself. The probability that a word beginning WB is a morph, is defined as the ratio of the number of distinct letters $L(\text{WB})$ which can follow WB over the total number of distinct letters $L$. The division by $L$ is not mandatory since it is a constant and thus does not influence the cost minimization, but it was kept for coherence, since the other quantities used in the algorithm are probabilities. This term is inspired from Harris' observation that this number decreases naturally from the word start, and that if it increases at a given point in the string, the sub-string up to this point might be a morph, that can be followed by many different suffixes

$$P_H(\text{WB}) = \frac{L(\text{WB})}{L}. \quad (4)$$

This definition favors short morphs, which is potentially interesting for languages where the word compounding generation process corresponds to the addition of prefixes and suffixes that are grammatical morphemes such as pronouns, possessive and demonstrative adjectives, prepositions, and postpositions.

### C. Modified Algorithm for ASR

All the properties used in the Morfessor program are based on written language and do not incorporate any "oral" properties that could be useful for ASR. Two modifications were introduced to try incorporate such properties.

*1) Distinctive Feature Motivated Property:* This property is an attempt to incorporate linguistic knowledge in the decompounding process. A phone-based feature was added to the $P(L)$ term of (2) and (3). This property aims to give an estimation of the phonemic confusability between lexical units. It is theoretical and relies on some distinctive features (DF) of the phones used in the language of study. The DFs are basically the same as those used in the decision tree that merges contexts during acoustic model training (as described in the experimental Section V-D). For a particular morph, the smaller the feature value is, the greater the number of similar morphs (in terms of DFs) there are in the lexicon. As for the other terms of the Morfessor algorithm, it takes the form of a probability.

Equation (5) gives the definition for a morph $m_k$. The DFs of its vowels are compared to the DFs of the vowels of all the other morphs that share the same consonantal root. The compared vowels have the same position in the morphs being compared. The same definition is used for consonants, however in that case, the DFs of morphs that share the same "vocal root" are compared. For example, the two Amharic words with the phoneme transcriptions of [nɛwa], [nɛwə], share the same $[n, w]$ consonantal root. Thus the vowel DFs are compared. Both words have the same first vowel, which is ignored in the computation, otherwise the feature would be zero. Only the vowel pair [a,ə] will have a contribution. The other possible vowel pairs [ɛ,a] and [ɛ,ə] are not used since they involve vowels that have different word positions. In an analogous manner, if two words share the same "vocal root," then DF differences in the consonants can be computed.

Two distinct results, one for the vowels of morph $m_k$ and one for its consonants can be computed. In the next sections, results using only the vowel DFs, only the consonant DFs, and both DFs (computed by summing their logarithms) will be given.

The following discussion explains how this feature is computed for vowels, the extension to consonants being straightforward. Equations (5) and (6) are used to define how the difference in score of distinctive features for vowels is computed as follows:

$$P_{DF}(m_k) = \prod_{j=1, j\neq k}^{j=N_k} P_{DF}(m_k, m_j) \quad (5)$$

TABLE III
DISTINCTIVE FEATURES OF THE AMHARIC VOWELS, USED WITH THE ALGORITHM. REMARK: BASED ON THE VOWEL CONFUSIONS REPORTED IN A PREVIOUS STUDY [10], FOR THIS STUDY THE /A/ IS CONSIDERED NON-TENSE

| *DF* | *Vowels* | | | | | | |
|---|---|---|---|---|---|---|---|
| *IPA* | ɛ | u | i | a | e | ə | ɔ |
| *high* | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| *low* | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| *round* | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| *tense* | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| *reduced* | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| *back* | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| *long* | 0 | 1 | 1 | 1 | 1 | 0 | 1 |

with

$$P_{DF}(m_k, m_j) = \prod_{l=1}^{l=V_k} \frac{\Delta_{kl,jl}}{C} \tag{6}$$

where $N_k$ is the number of morphs that share the same consonantal root, $\Delta_{kl,jl}$ is the number of DFs in which the $l$th vowels of morphs $m_k$ and $m_j$ differ (computed only if the vowels are different), $V_k$ is the total number of vowels in morph $m_k$, and $C$ is the total number of distinct DFs. Note that while $P_{DF}(m_k) \in [0, 1]$, $P_{DF}$ is not a probability since it does not sum to one. The more distinct DFs two morphs have, the bigger the feature value is, and the smaller the associated "cost" (negative logarithm of $P_{DF}$) is. This feature thus aims to favor word decompositions that give morphs which have distinct DFs compared to the other morphs.

To evaluate $\Delta_{kl,jl}$, one can use standard DF tables found in phonetics literature, for example in [40]. The distinctive features used in this study concern vowels and consonants, and are given for information in Table III for vowels only. Features for consonants are similar ([41, p. 144]).

Finally, as shown in (7), $P_{DF}$ has been incorporated in $P(L)$ as an additional term. Equation (7) is our modified version of the original $P(L)$ Morfessor formulation, given in (3). As for the other three properties ($n$, $s$, $l$), the property $P_{DF}(m_k)$ is considered to be independent from the other morph feature values so that $P_{DF}(m_1, \ldots, m_M) = \prod_{k=1}^{M} P_{DF}(m_k)$.

$$\begin{aligned} P(L) = P(n(m_1), \ldots, n(m_M))P(s(m_1), \ldots, \\ (m_M))P_H(l(m_1), \ldots, l(m_M))P_{DF}(m_1, \ldots, m_M). \end{aligned} \tag{7}$$

*2) Phonemic Confusion Constraint:* The DF property is theoretical and therefore does not account for the phonological variation observed in real world speech, such as in the choice of vowel alternatives. In [12], syllabotactic alignments were studied in order to determine the most frequent confusions at the syllable level. For each syllable, the vowel that was most often substituted by the aligner was determined. These confusion pairs provide an additional means of reducing phonemic confusion amongst units arising from the decompounding.

During the decompounding process, candidates for word splitting that differ from other morphs by only one syllable are compared. If the pair of syllables is among the most frequently confused pairs found in the alignment study, the candidate is rejected (the split is refused). In the previous example with the two words [nɛwa], [nɛwə],, if the algorithm already split the first word into [n$\epsilon$ + wa], and if the split of the second word into [nɛ+ wə] was found to lower the global function cost and thus be a good decomposition, the Cc constraint would forbid this decomposition if the syllable pair [wa] and [wə] was among the confusion pairs resulting from the syllabotactic alignments.

TABLE IV
DECOMPOSITION OPTIONS COMPARED IN THIS STUDY

| *Option* | *Comment* |
|---|---|
| BL | Baseline word based system, no decompounding |
| M | Baseline Morfessor 1.0 |
| M H | M + modified 'Harris' |
| M H DFV | M H + distinctive features parameter of vowels only |
| M H DFC | M H + distinctive features parameter of consonants only |
| M H DFCV | M H + distinctive features parameter of vowels and consonants |
| Cc | + confusion constraint |

The different options investigated with the decompounding algorithm are summarized in Table IV. The configurations M, MH, MHDFV, MHDFC, and MHDFCV are compared both with and without the confusion constraint Cc.

## V. EXPERIMENTAL STUDY

In this section, recognition experiments for the Amharic language are reported using a corpus of broadcast news data.

### A. Amharic Corpus

Some recent studies, for example [9], [42], have addressed speech recognition and speech processing for Amharic using read speech. In the experiments reported here, a broadcast news speech corpus is used. Compared to other languages for which models and systems have been developed [43], the available Amharic audio corpus is quite small. It is comprised of 37 h of broadcast news data from two sources, *Deutsche Welle* (25 h 26 min) and *Radio Medhin* (11 h 45 min). The data were transcribed by native Ethiopian speakers, and contain a total of 247k words with 50k distinct lexemes. Two hours of data taken from the latest shows of each source were reserved for development and test. This data contains 14.2 k words, of which almost 15% do not appear in the training portion. In a previous study [14], results were reported on the same 2-h corpus that was used for development purposes. This means that certain parameters, such as the language model interpolation coefficients were optimized on the data potentially introducing a bias. Since no additional data are available, for the experiments reported in this paper, the same 2-h corpus was divided into two distinct subsets, 80% for development, and 20% for test (percentages based on the number of words). Seven distinct dev/test configurations were randomly selected, in order to do a classical cross-validation. Table V gives the number of speakers and words in the different subsets. Depending on the randomly selected files for dev/test, the number of speakers is between 12 and 15 for the dev, 4 and 7 for test.

In addition to the transcriptions of the audio data, about 4.6 million words of newspaper and web texts have been used for language model training. Over 340 k distinct words are found in these texts.

TABLE V
CHARACTERISTICS OF THE AUDIO CORPUS (NUMBER OF HOURS, SPEAKERS, AND TOTAL NUMBER OF WORDS FOR TRAINING, DEV AND TEST)

| | *Training* | *Development* | Test |
|---|---|---|---|
| | 35hr 14min | 1hr 34min | 23min |
| *# Speakers* | 200 | [12-15] | [4-7] |
| *# Words* | 233k | 11.4k | 2.8k |

TABLE VI
NUMBER OF MORPH TYPES IN THE LEXICONS WITH AND WITHOUT "+" FOR DIFFERENT DECOMPOUNDING OPTIONS. (BL: WORD-BASED SYSTEM, M: BASELINE MORFESSOR, H: HARRIS' OPTION, CC: CONFUSION CONSTRAINT, DFV: DISTINCTIVE FEATURES FOR VOWELS, DFC: DISTINCTIVE FEATURES FOR CONSONANTS, DFCV: DISTINCTIVE FEATURES FOR VOWELS AND CONSONANTS)

| *Options* | Lexicon size | |
|---|---|---|
| | with '+' | without '+' |
| BL | 0 | 133384 |
| M | 95937 | 70267 |
| M Cc | 128239 | 109694 |
| M H | 90740 | 65421 |
| M H Cc | 126105 | 107123 |
| M H DFV | 94198 | 69038 |
| M H DFV Cc | 128404 | 110320 |
| M H DFC | 66190 | 50062 |
| M H DFC Cc | 118596 | 101770 |
| M H DFCV | 66250 | 50193 |
| M H DFCV Cc | 107786 | 93573 |

### B. *Decompounding the Training Texts*

When building a recognition lexicon from training texts, a frequency cutoff is typically applied to get rid of misspelled words and artifacts. In this study the cutoff is applied after decomposition. It should be noted that given the CV structure of the Amharic language, word splits are allowed only after a vowel. First, a decompounding model is built for a reference lexicon, and then this model is used to decompose all words in the corpus without any frequency cutoff. A new reference lexicon is then selected, applying a frequency cutoff: only morphs occurring at least three times are included in the lexicon. The OOV rate may decrease since OOV words may have been decompounded. The number of lexical tokens in the training text corpus is also increased with this method.

An initial 133 k word-based lexicon was selected. It was comprised of the 50 k distinct words in the acoustic training data transcriptions and all words occurring at least three times in the newspaper and web texts. The out-of-vocabulary rate of the development corpus with this word list is almost 7%, which is quite high compared to the OOV rates obtained for well-represented languages which are typically around 1%.

Table VI shows the number of morph types for the different decompounding options listed in Table IV. Since a morph may exist both as a word and as an affix, the explicit use of this information is investigated by adding a "+" sign to prefixes found by the algorithm in order to simplify the work of recombining morphs back into entire words in the ASR experiments. The distinctive feature option for consonants (DFC) gives the smallest lexicon with about 66 k units, being about half the size of the original lexicon. The Cc constraint increases lexicon size by 25%–30% relative to the same configuration without the constraint, except for the DFC option, for which the increase is quite a bit larger (43%). This indicates that the use of DFC splits many words into potentially confusable sub-word units. Since the word and affix entries corresponding to the morph will have the same pronunciations in the recognition lexicon, the choice between forms is made by the language model. The third column gives for information the number of types when no explicit distinction is made between words and affixes (i.e., no "+" sign is added during decomposition). The difference between the second and the third columns is the number of morphs that are also words.

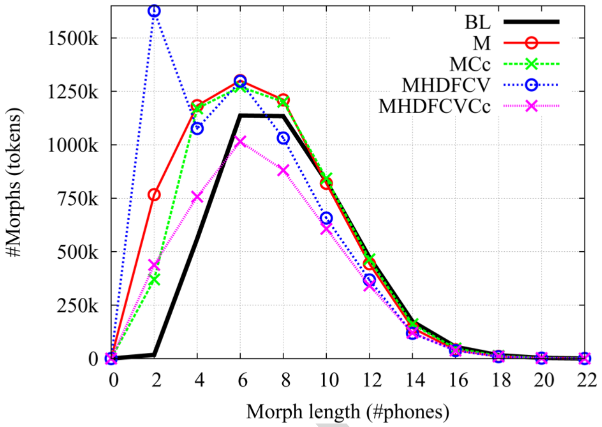

Fig. 1. Number of morph tokens in the training data as a function of the number of phones for different decomposition options. (BL: word-based system, M: baseline Morfessor, Cc: confusion constraint, DFCV: distinctive features for vowels and consonants).

Fig. 1 shows the number of tokens as a function of their length in phones,[2] for different decompounding options. The BL curve (in black) is the baseline curve, with no decompounding. The other curves, for which words were decompounded, show a noticeable shift to smaller word lengths. Some decompounding options have been omitted to keep the figure readable, but these curves are similar to ones shown. The curves with and without the "Cc" option form two distinct groups. As expected, the "non Cc" curves (drawn with "o" points) have substantially more morph tokens with a length of 2 phones compared to the "Cc" curves (drawn with "x" points), since more words are decompounded without the constraint. Basically, the DF property for consonants (DFC) introduces the largest number of small units, and the M H DFCV curve have almost twice as many 2-phone units than the other "non Cc" curves. As was written in the introduction, small units are more error-prone than longer units (see [10], [11]). Reducing their frequency with the phonetic "Cc" constraint is thus promising, but of course results in a larger lexicon size and/or OOV rate.

### C. *Language Model and OOV Rates*

The language models are Kneser–Ney smoothed four-gram models, and result from the interpolation of two component LMs: one estimated on the web/newspaper texts and the other on the manual transcripts of the audio training data. The interpolation coefficient was optimized for each LM by measuring the

[2] Recall that characters in Amharic correspond to a syllable, so all points are multiples of 2 phones since the lengths are determined from a canonical pronunciation.

TABLE VII
AVERAGED OOV RATES (%) ON THE TEST CORPUS. (BL: WORD-BASED SYSTEM, M: BASELINE MORFESSOR, H: HARRIS' OPTION, CC: CONFUSION CONSTRAINT, DFV: DISTINCTIVE FEATURES FOR VOWELS, DFC: DISTINCTIVE FEATURES FOR CONSONANTS, DFCV: DISTINCTIVE FEATURES FOR VOWELS AND CONSONANTS)

| *Options* | *OOV Tokens (%)* |
|---|---|
| BL | 6.83 |
| M | 3.99 |
| M Cc | 4.32 |
| M H | 3.97 |
| M H Cc | 4.30 |
| M H DFV | 3.97 |
| M H DFV Cc | 4.35 |
| M H DFC | 3.47 |
| M H DFC Cc | 4.10 |
| M H DFCV | 3.47 |
| M H DFCV Cc | 4.35 |

TABLE VIII
WORD ERROR RATES FOR THE DIFFERENT ASR SYSTEMS. (BL: WORD-BASED SYSTEM, M: BASELINE MORFESSOR, H: HARRIS' OPTION, CC: CONFUSION CONSTRAINT, DFV: DISTINCTIVE FEATURES FOR VOWELS, DFC: DISTINCTIVE FEATURES FOR CONSONANTS, DFCV: DISTINCTIVE FEATURES FOR VOWELS AND CONSONANTS). OOV RATES WITH THE INITIAL 133 K LEXICON ARE ALSO GIVEN FOR EACH BATCH

| | *WER (%)* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Algorithm* | *Batch* | | | | | | | |
| *Options* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | *Mean* |
| OOV | 7.3 | 5.6 | 8.7 | 6.7 | 5.6 | 6.7 | 7.4 | 6.8 |
| BL | 23.8 | 19.8 | 24.1 | 24.3 | 23.2 | 22.3 | 27.4 | 23.6 |
| M | 24.8 | 20.1 | 23.4 | 24.8 | 23.4 | 22.6 | 28.2 | 23.9 |
| M Cc | 23.4 | 19.6 | 22.7 | 23.8 | 22.5 | 21.3 | 27.5 | 23.0 |
| M H | 24.9 | 20.2 | 23.3 | 24.7 | 23.5 | 23.0 | 28.5 | 24.0 |
| M H Cc | 23.9 | 19.3 | 23.2 | 23.8 | 22.3 | 21.3 | 27.0 | 23.0 |
| M H DFV | 24.6 | 20.2 | 22.9 | 24.6 | 23.4 | 22.4 | 28.8 | 23.8 |
| M H DFV Cc | 23.5 | 19.6 | 23.0 | 23.6 | 22.5 | 21.5 | 27.2 | 22.9 |
| M H DFC | 24.9 | 20.0 | 23.5 | 25.0 | 23.9 | 22.2 | 27.5 | 23.8 |
| M H DFC Cc | 24.0 | 19.5 | 22.7 | 23.7 | 22.5 | 21.0 | 27.2 | 22.9 |
| M H DFCV | 24.8 | 20.3 | 23.4 | 25.3 | 24.0 | 22.2 | 28.9 | 24.1 |
| M H DFCV Cc | 23.6 | 19.9 | 22.9 | 24.3 | 22.9 | 21.3 | 27.0 | 23.1 |

perplexity on the development transcripts. Different LMs were built for each set of decompounding options, and for each development/test subdivision. Since some of the words which are not in the baseline vocabulary are decomposed, the OOV rates are reduced. Table VII gives the mean token OOV rates averaged across the seven different test subsets (each about 2.8 k words). The relative reduction in OOV rate ranges from 35% to 50% depending on the options.

### D. ASR Experiments

This section reports recognition results obtained with systems trained for each of the decomposition option configurations. The baseline system is the word-based system. The speech recognizers all have two decoding passes, with unsupervised acoustic model adaptation (MLLR) after the first decoding pass [44]. The acoustic models are all tied-state triphone HMMs, covering both word-internal and cross-word contexts, with three states per model and 32 Gaussians per state. State tying is based on classical decision tree clustering, with backoff on diphones and monophones. The set of questions concern the phone position, the distinctive features (and identities) of the phone and the neighboring phones. Since different decompositions result in different recognition units (and therefore different word positions), it was necessary to build specific acoustic models for each set of options. In all cases both intra- and cross-recognition unit contexts are modeled. All acoustic model sets cover about 10.5 k distinct contexts, with a total of about 8.5 k tied states.

Table VIII gives the OOV and word error rates (WER) for the different ASR systems, for the seven development/test configurations, estimated after recombining prefixes (that end with a "+" sign) and roots back into full words. The means of the WERs over the seven configurations are given in the last column. The OOV rate for the word-based system ranges from 5.6% to 8.7%, with an average of 6.8%, which is close to that of the full development data set (6.9%) used in [14]. The largest OOV rate is for subset 3, and the smallest rates are for subsets 2 and 5. The full-word baseline system has a mean WER of 23.6%. The five systems M, MH and MHDFV, MHDFC, MHDFCV, which do not use the confusion constraint Cc, perform slightly less well than the baseline system. On the contrary, the five Cc systems all give small gains. The confusion constraints between lexical units appears to be useful for identifying recognition units when used in conjunction with word decompounding. The worst performance is obtained by the MHDFCV system, which is the algorithm that splits the largest number of words. This result illustrates well the compromise between OOV rate reduction and increased confusability between lexical units when decompounding is used.

The Harris modification seems useful since it produces smaller lexicons than with Morfessor baseline, and the same mean WER is obtained when using the Cc option (23% WER for both MCC and MHCc systems). Concerning the DF option, there is a 0.4% absolute WER reduction between the MHDFV, MHDFC and MHDFCV systems and their corresponding Cc version. The best performance is obtained with the DFV and DFC motivated systems (MHDFVCc and MHDFCCc) which achieves a 0.7% absolute improvement compared to the baseline. Nevertheless, the WERs of the two systems vary depending on the dev/test subdivision, which can surely be attributed to the small size of the individual sets. It can be seen that results on batch number 3 are different from the other batches in that all the morph-based systems performed better than the word-based system. This may be due to the higher OOV rate of this subset (8.7%) with the baseline system. Significance tests at word-level (MAPSSWE) have been conducted with the "sc_stats" NIST tool, available at www.nist.gov/speech/tools. In comparison with the word-based system, the system based on the baseline Morfessor algorithm does not show any significative difference for any of the batches, although it performs slightly worse on all the test sets, with the exception of batch 3. The two best systems (MHDFVCc and MHDFCCc) show significative differences in performance with the classical 95% confidence threshold, only for batch number 3. For test sets 4, 5, and 6, the threshold is about 85%, and for the others, the performance difference is not significative. This indicates that the modifications seem more useful with test sets that present the highest OOV rates.

By comparing the distinct types of errors, with the percentages of insertions, deletions and substitutions, it appears that

TABLE IX
COMPARISON OF THE WORD ERROR RATES (WER) FOR THE MHDFV AND MHDFVCC SYSTEMS WITH FINAL 5-GRAM LM RESCORING WITH THE WER OF THE CORRESPONDING PREVIOUSLY USED SYSTEMS (4-GRAM LM). OOV RATES WITH WORD-BASED SYSTEMS AND PREVIOUS WERS OBTAINED WITH 4-GRAM RESCORING ARE ALSO GIVEN. IMPROVEMENTS (WER REDUCTIONS) ARE SHOWN IN BLUE AND DEGRADATIONS (WER INCREASES) ARE SHOWN IN RED (M: BASELINE MORFESSOR, H: HARRIS' OPTION, CC: CONFUSION CONSTRAINT, DFV: DISTINCTIVE FEATURES FOR VOWELS)

| *Algorithm* | *WER(%)* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Batch* | | | | | | | |
| *Options* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | *Mean* |
| OOV (%) | 7.3 | 5.6 | 8.7 | 6.7 | 5.6 | 6.7 | 7.4 | 6.8 |
| M H DFV 4g | 24.6 | 20.2 | 22.9 | 24.6 | 23.4 | 22.4 | 28.8 | 23.8 |
| M H DFV 5g | 24.3 / -0.3 | 20.3 / +0.1 | 22.9 / 0.0 | 24.4 / -0.2 | 23.3 / -0.1 | 22.3 / -0.1 | 28.5 / -0.3 | 23.7 / -0.13 |
| M H DFV Cc 4g | 23.5 | 19.6 | 23.0 | 23.6 | 22.5 | 21.5 | 27.2 | 22.9 |
| M H DFV Cc 5g | 23.7 / +0.2 | 19.4 / -0.2 | 23.0 / 0.0 | 23.7 / +0.1 | 22.4 / -0.1 | 21.3 / -0.2 | 27.1 / -0.1 | 22.9 / -0.04 |

all the morph-based systems have higher average deletion rates than the baseline (2.2% for the BL system versus 2.8% for the M system for example), but lower insertion rates (2.4% for the BL system, 2.1% for the M system). Systems which do not use the Cc constraint have higher substitution rates, suggesting that the Cc constraint is doing what it was designed to do. Looking at the decoder output, the systems (without Cc) do have a tendency to insert small morphs. However this effect is lost after recombining morphs into words. When the morphs are glued together, the errors are counted as substitutions when compared to the word based reference.

The two best systems (MHDFVCc and MHDFCCc) have similar insertion plus deletion rates as the baseline, but the substitution rate is a bit smaller (18.3% vs 19.0%). This improvement may be explained by the recognition of ex-OOV words, as analyzed in the next paragraph.

It was mentioned earlier that word decompounding possibly allows words that were OOV before decompounding to be recognized since sub-word units can be combined to form a word that was not in the initial lexicon. Using batch number 1 for analysis, the initial test OOV rate is 7.3% with 242 OOV tokens for a total of 3321 words. For all the sub-word based systems, about 80 of the OOV words are covered by the respective lexicon. Depending upon the system configuration 26 to 30 of these words were correctly recognized. For batch number 3, which has the highest OOV rate (8.7%), the number of words that are no longer OOV is larger (between 89 to 104 words depending on the system options). More than a half of these words were correctly recognized. For example, with the MH system, 55 ex-OOV words are correctly recognized. There are 2708 words in the associated reference transcripts for this batch, which would suggest that an absolute gain of 2.0% should have been observed. However, as can be seen in Table VIII, the gain is lower, only 0.8% absolute, therefore new errors, i.e., some that were not produced by the word-based system, are introduced by the use of the sub-word units, increasing the error rate by 1.2% absolute. Additional errors may be due to ungrammatical morph sequences, corresponding to the phenomenon called "over-generation." For batch number one for example, the M system output 116 words that are not in the baseline word-based lexicon. 27 of these words were correctly recognized, and correspond to some ex-OOV words as explained above. The remaining 89 words are possibly the result of over-generation, allowing an upper limit on the errors due to over-generation to be estimated at 2.7%. For the MCc system, that creates less decompositions, this estimated upper limit is lower (1.9%). Looking at some of these words, it is clear that these values overestimate the number of introduced errors, since the great majority of these words were already misrecognized with the baseline system. Finally, considering the very permissive rules of Amharic orthography, only an Amharic expert can identify ungrammatical morph sequences properly [45].

When sub-word units are used, the effective span of an n-gram language model is reduced. Shorter units naturally require longer n-grams. In [4] for example, speech recognition experiments were carried out with 5-gram, 7-gram, and 8-gram LMs for respectively Turkish, Finnish and Estonian. The results reported previously in this section all used a 4-gram span for the language models, which may favor the word-based system. A complementary experiment with 5-gram LMs has been conducted.

In this experiment, a rescoring with 5-gram LMs of the best hypotheses is carried out, using the lattices generated by the previously used decoders (these lattices were generated by 4-gram LMs). Results are reported for the MHDFV Cc system, that had the best recognition performance, and with the MHDFV system to evaluate the impact of the Cc option. Table IX gives the absolute differences between WERs obtained with 5-gram rescoring, in comparison to the previous WERs obtained with 4-gram rescoring, reminded in the table. The lattices used for rescoring are identical, the only change is in the order of the LMs. In the table, improvements are shown in blue and increases in WER are shown in red. Globally, for both options, there are more improvements than degradations in performance, but the differences are quite small. The average improvement for the MHDFV system is larger than for the corresponding system with the Cc option (0.13% against 0.04% in mean), which seems natural since there are more word decompositions without the Cc constraint, and therefore more small morphs that can benefit from a longer-span LM. However for both systems, the differences are very small, and further experiments with longer-span LMs would be needed to draw firm conclusions.

## VI. DISCUSSION

In this paper, sub-word units have been investigated to address the issue of very large lexical variety found in morphologically-rich languages, for the task of automatic speech recognition of broadcast news data. An unsupervised data-driven word decompounding algorithm, which extends the Morfessor algorithm to better suit speech recognition, has been described. The original and modified algorithms have been tested in recognition experiments, where OOV and WER reductions have been obtained on a morphologically rich and less-represented language in which grammatical morphemes are glued to roots. For some systems, gains in performance were achieved since words that were out-of-vocabulary with respect to an initial word lexicon were able to be recognized by the morphologically decompounded ones. Nevertheless, the use of small units, often referred to as "morphs," introduces new errors due to an increased confusability between lexical units. This article has attempted to address both the problem of high OOV rates observed for morphologically rich languages, and the problem of increased confusability when using sub-word units as recognition units.

The "Morfessor" algorithm splits words into smaller units in an iterative manner by maximizing a MAP estimate of a lexicon given a word list with frequency counts. The end-of-word probability computation has been modified to allow more splits. A new phonemic-based parameter motivated by distinctive features (DF), was incorporated as well as phonemic confusion constraints derived from previous experiments with automatic alignment of audio data. Systems built with different combinations of options were compared. The best systems all include the confusion constraint, and the phonemic-based DF parameter for consonants or for vowels. For these systems, the lexicon sizes are reduced by about 10%, along with a small absolute gain in WER (0.7%) relative to the reference error (23.6%) of the word-based system. Without the confusion constraints, all systems had slightly worse performance than the baseline. This result demonstrates the usefulness of the confusion constraints. Without the constraint, small units (2 phones long) are very frequent and somewhat error-prone. The differences in errors of the word based and sub-word based systems were analyzed in order to assess how successful the approach is recovering errors due to words that were OOV for the word-based system. The recognition of "ex-OOV" words gives around a 2% absolute gain, but since new errors are introduced, the overall gain is smaller. Contrastive experiments with longer-span LMs (5-gram LMs) were conducted, but showed very little improvement over the 4-gram LMs used throughout this work.

The DF parameter is a phonetically motivated parameter, introduced for vowels and for consonants. Further investigation should be carried out to confirm the usefulness of this parameter. In the current implementation, the different terms in the MAP estimate are summed, however it may be useful to weight these terms in order to optimize each contribution. Future plans are to test the algorithm on another language similar to Amharic, Arabic for instance, for which ample training data are available, as well as on a language in which the word compounding generation process is even more important, such as German or Turkish.

## References

[1] J.-L. Gauvain, L. Lamel, G. Adda, and M. Adda-Decker, "Speaker-independent continuous speech dictation," in *Speech Commun.*, 1994, vol. 15, pp. 21–37.

[2] M. Adda-Decker, "A corpus-based decompounding algorithm for German lexical modeling in LVCSR," in *Proc. Eurospeech*, Geneva, 2003, pp. 257–260.

[3] P. Geutner, "Using morphology towards better large-vocabulary speech recognition systems," in *Proc. ICASSP*, Detroit, 1995, pp. 445–448.

[4] ***[AUTHOR: Please provide page range]***M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and M. Saraclar, "Unlimited vocabulary speech recognition for agglutinative languages," in *Proc. HLT-NAACL*, New York, 2006.

[5] ***[AUTHOR: Please provide page range]***V.-B. Le, L. Besacier, S. Seng, B. Bigi, and T.-N.-D. Do, "Recent advances in automatic speech recognition for vietnamese," in *Proc. SLTU*, Hanoi, Vietman, 2008.

[6] R. Ordelman, A. van Hessen, and F. de Jong, "Compound decomposition in Dutch large vocabulary speech recognition," in *Proc. Eurospeech*, Geneva, 2003, pp. 225–228.

[7] M. Creutz and K. Lagus, "Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0," Computer and Information Science, Report A81, 2005.

[8] T. Schultz, A. Black, S. Badaskar, M. Hornyak, and J. Kominek, "SPICE: Web-based tools for rapid language adaptation in speech processing systems," in *Proc. Interspeech*, Antwerp, 2007, pp. 2125–2128.

[9] S. Abate and W. Menzel, "Automatic speech recognition for an under-resourced language—Amharic," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 1541–1544.

[10] T. Pellegrini and L. Lamel, "Investigating automatic decomposition for ASR in less represented languages," in *Proc. Interspeech*, Pittsburgh, PA, 2006, pp. 285–288.

[11] K. Kirchhoff, J. Bilmes, J. Henderson, and R. Schwartz, "Novel speech recognition models for Arabic," in *Johns Hopkins University Summer Research Workshop*, 2002, Final report.

[12] T. Pellegrini and L. Lamel, "Experimental detection of vowel pronunciation variants in Amharic," in *Proc. LREC*, Genoa, Italy, 2006, pp. 1005–1008.

[13] R. Jakobson, G. Fant, and M. Halle*, Preliminaries to Speech Analysis*. Cambridge, MA: MIT Press, 1952.

[14] T. Pellegrini and L. Lamel, "Using phonetic features in unsupervised word decompounding for asr with application to a less-represented language," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 1797–1800.

[15] A. Martinet*, Éléments de Linguistique Générale*. Paris, France: Armand Colin, 1980.

[16] G. Adda, M. Adda-Decker, J.-L. Gauvain, and L. Lamel, "Text normalization and speech recognition in French," in *Proc. EuroSpeech*, Rhodes, Greece, 1997, vol. 5, pp. 2711–2714.

[17] F. Van Eynde and D. Gibbon*, Lexicon Development for Speech and Language Processing*. Dordrecht, The Netherlands: Kluwer, 2000.

[18] M. Finke and A. Waibel, "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition," in *Proc. Eurospeech*, Rhodes, 1997, pp. 1963–1966.

[19] J.-L. Gauvain, G. Adda, L. Lamel, and M. Adda-Decker, "Transcribing broadcast news: The Limsi Nov96 Hub4 System," in *Proc. ARPA*, Chantilly, France, 1997, pp. 56–63.

[20] ***[AUTHOR: Please provide page range]***A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. G. Rao, M. Plauche, C. Richey, E. Shriberg, K. Sonmez, F. Weng, and J. Zheng, "The SRI March 2000 hub-5 conversational speech transcription system," in *Proc. NIST Speech Transcription Workshop*, College Park, MD, 2000.

[21] ***[AUTHOR: Please provide page range]***T. Emerson, "The second international chinese word segmentation bakeoff," in *Proc. 4th SIGHAN Workshop Chinese Lang. Process.*, Jeju, Korea, 2005.

[22] L. Chen, L. Lamel, and J.-L. Gauvain, "Transcribing mandarin broadcast news," in *Proc. IEEE Workshop Autom. Speech Recognition*, St. Thomas, Virgin Islands, 2003, pp. 99–104.

[23] M.-Y. Hwang, X. Lei, W. Wang, and T. Shinozaki, "Investigation on Mandarin broadcast news speech recognition," in *Proc. ICSLP*, Pittsburgh, PA, 2006, pp. 1233–1236.

[24] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraclar, and A. Stolcke, "Morph-based speech recognition and modeling of out-of-vocabulary words across languages," *ACM Trans. Speech Lang. Process.*, vol. 5.1 Article 3, 2007.

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

[26] E. Arisoy, H. Sak, and M. Saraclar, "Language modeling for automatic Turkish broadcast news transcription," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 2381–2384.

[27] E. Arisoy and M. Saraclar, "Lattice extension and vocabulary adaptation for Turkish LVCSR," *Speech Lang. Process.*, vol. 10, no. 1, 2009.

[28] M. Afify, R. Sarikaya, H.-K. Kuo, L. Besacier, and Y. Gao, "On the use of morphological analysis for dialectal Arabic speech recognition," in *Proc. ICSLP*, Pittsburgh, PA, 2006, pp. 277–280.

[29] B. Xiang, K. Nguyen, L. Nguyen, R. Schwartz, and J. Makhoul, "Morphological decomposition for Arabic broadcast news transcription," in *Proc. ICASSP*, Toulouse, France, 2006, vol. I, pp. 1089–1092.

[30] L. Lamel, A. Messaoudi, and J.-L. Gauvain, "Investigating morphological decomposition for transcription of Arabic broadcast news and broadcast conversation data," in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 1429–1432.

[31] Omniglot, "A guide to the languages, alphabets, syllabaries and other writing systems of the world." [Online]. Available: http://www.omniglot.com/

[32] L. Asker, A. Argaw, B. Gambäck, and M. Sahlgren, "Applying machine learning to Amharic text classification," in *Proc. 5th World Congr. African Linguist.*, Cologne, Germany, 2007.

[33] D. Appleyard, *Colloquial Amharic*. London, U.K.: Routledge, 1995.

[34] ***[AUTHOR: Please provide page range]***K. Kirchhoff and R. Sarikaya, "Processing morphologically rich languages," in *Proc. Workshop Interspeech*, Antwerp, Belgium, 2007.

[35] P. Geutner, M. Finke, and A. Waibel, "Phonetic-distance-based hypothesis driven lexical adaptation for transcribing multilingual broadcast news," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 771–774.

[36] P. Geutner, C. Carki, and T. Schultz, "Turkish LVCSR: Towards better speech recognition for agglutinative languages," in *Proc. ICASSP*, Istanbul, Turkey, 2000, pp. 1563–1566.

[37] E. Arisoy, H. Dutagaci, and L. Arslan, "A unified language model for large vocabulary continuous speech recognition of Turkish," *Signal Process.*, vol. 86, no. 10, pp. 2844–2862, 2006.

[38] Z. Harris, "From phoneme to morpheme," in *Language*, 1955, vol. 31, pp. 190–222.

[39] J. Goldsmith, "Unsupervised learning of the morphology of a natural language," *Comput. linguist.*, vol. 27, no. 2, pp. 153–198, 2001.

[40] M. Halle and G. Clements, *Problem Book in Phonology*. Cambridge, MA: MIT Press, 1983.

[41] T. Pellegrini, "Transcription Automatique de Langues peu Dotées" Ph.D., Paris-Sud Univ., Orsay, 2008.

[42] S. Abate, W. Menzel, and B. Tafila, "An Amharic speech corpus for large vocabulary continuous speech recognition," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 1601–1604.

[43] L. Lamel, J.-L. Gauvain, G. Adda, M. Adda-Decker, L. Canseco, L. Chen, O. Galibert, A. Messaoudi, and H. Schwenk, "Speech transcription in multiple languages," in *Proc. ICASSP*, Montreal, QC, Canada, 2004, vol. 3, pp. 757–760.

[44] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Commun.*, vol. 37, no. 1–2, pp. 89–108, 2002.

[45] ***[AUTHOR: Please provide page range]***D. Yacob, "Application of the double metaphone algorithm to Amharic orthography," in *Proc. Int. Conf. Ethiopian Studies XV*, Cologne, Germany, 2003.

**Thomas Pellegrini** graduated with a degree in physics from the Ecole Supérieure de Physique et de Chimie Industrielles (ESPCI), Paris, France, in 2003 and received the M.Sc. degree in acoustics, signal processing, computer science applied to music from IRCAM, Paris VI University (ATIAM), in 2003 and the Ph.D. degree in computer science on speech recognition for less-represented languages from LIMSI-CNRS from the Paris-Sud University, in 2008.

He was a Teaching Assistant at Paris La Sorbonne from 2007–2008, teaching mainly object-oriented programming. Since September 2008, he has been a Postdoctoral Researcher at the Spoken Language Systems Lab (L$^2$F), Lisbon, Portugal. His research interests are automatic speech recognition, audio, music, and sound in general.

**Lori Lamel** (M'88) received the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, in 1988.

She is a Senior CNRS Researcher in the Spoken Language Processing group, LIMSI, Orsay, France, which she joined in October 1991. Her principal research activities are in speech recognition, studies in acoustic–phonetics, lexical and phonological modeling, and speaker and language identification. She has been a prime contributor to the LIMSI participations in DARPA benchmark evaluations and developed the American English pronunciation lexicon. She has been involved in many European projects, most recently the IPs Chil, and TCStar. She has over 200 reviewed publications.

Dr. Lamel is a member of the Speech Communication Editorial Board and the Interspeech International Advisory Council. She was a member of the IEEE Signal Processing Society's Speech Technical Committee from 1994 to 1998, and the Advisory Committee of the AFCP, the IEEE James L. Flanagan Speech and Audio Processing Award Committee (2006–2009) and the EU-NSF Working Group for "Spoken-Word Digital Audio Collections." She is a corecipient of the 2004 ISCA Best Paper Award for a paper in the *Speech Communication Journal*.

# Automatic Word Decompounding for ASR in a Morphologically Rich Language: Application to Amharic

Thomas Pellegrini and Lori Lamel, *Member, IEEE*

***Abstract*—This paper investigates a data-driven word decompounding algorithm for use in automatic speech recognition. An existing algorithm, called "Morfessor," has been enhanced in order to address the problem of increased phonetic confusability arising from word decompounding by incorporating phonetic properties and some constraints on recognition units derived from forced alignments experiments. Speech recognition experiments have been carried out on a broadcast news task for the Amharic language to validate the approach. The out of vocabulary (OOV) word rates were reduced by 35% to 50% and a small reduction in word error rate (WER) has been achieved. The algorithm is relatively language independent and requires minimal adaptation to be applied to other languages.**

***Index Terms*—Automatic speech recognition (ASR), broadcast news transcription, less-represented languages, lexical modeling, morphologically rich languages (MRLs).**

## I. Introduction

In the literature, languages such as Arabic, Finnish, Turkish, and Estonian, are often referred to as "morphologically rich languages" (MRLs). Other languages do not have a "poor" morphology, this qualification emphasizes the highly productive processes involved in word formation in MRLs. For such languages, it is common to generate words by the compounding of smaller units that are primarily lexical morphemes (such as in German), or mostly grammatical morphemes (for example, Semitic languages such as Arabic or Amharic), or both (such as Turkish). These languages need very large lexicons, containing several hundred thousand words, to achieve good lexical coverage. Since state-of-the-art automatic speech recognition (ASR) systems generally use fixed (also called closed) lexicons, only the words in the recognition lexicon can potentially be recognized. For MRLs, the rich morphology implies a high number of unknown or out-of-vocabulary (OOV) words, which typically produce 1.5 to 2 errors for each OOV word [1]. This large lexical variety also poses a problem for language modeling, where it can be difficult to have reliable n-gram estimates for infrequent words. To address these issues, word decomposition has been investigated in a number of studies for various languages such as German [2], [3], Turkish, Finnish and Estonian [4], Vietnamese [5], and Dutch [6]. The probabilistic word decomposition framework used in this study is derived from the baseline version of the corpus-based word decompounding algorithm "Morfessor" [7].

High OOV rates and poor language model estimation are problems that also arise when developing technologies for less-represented languages, for which little data are available in an electronic form. Most of the world's languages suffer from poor representation on the web, which is being used more and more as the primary source for collecting data (principally texts) for building ASR systems [8]. This study reports on experiments carried out with the Amharic language, the official language of Ethiopia, which is both a less-represented language and a language in which grammatical compounding is frequent [9]. For morphologically rich languages and less-resourced languages, the first issue to be addressed, is the high percentage of unseen words as typical OOV rates are higher than 7%. Previous work reported improvements in ASR for Amharic broadcast news data when using sub-word units: for a relative OOV reduction of 16%, a 10% relative reduction in word error rate (WER) was achieved [10]. The sub-word units were identified with a character-based maximum branching factor algorithm similar to the one used in [2], and selected using a heuristic. In the same study, it was shown that experiments allowing more decompositions led to increased insertion and deletion rates, and to an overall degradation in performance (a 7% relative increase in WER, with a 20% relative OOV reduction compared to the best sub-word based system). A common observation in the literature is that small lexical units can often be less reliably decoded than longer units, since these units are acoustically more similar and therefore more confusion-prone [11]. One solution to overcome the increased confusion, consists of using word-based models to generate N-best lists or lattices, and a sub-word unit language model, only in a final rescoring framework. Nevertheless, for several reasons it was chosen to investigate the use of sub-word units in all stages of the decoder. First, Amharic is a language that has a rather straightforward grapheme-to-phoneme conversion, allowing pronunciations to be easily produced for sub-word units [12]. Second, the use of the same lexical units in all steps of the decoding simplifies the global process. Finally, we wanted to investigate new features that try to incorporate "oral" properties in the identification/selection of the sub-word units, in an attempt to take account of some specificities of spoken language. One of these new properties is based

Manuscript received August 04, 2008; revised April 10, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tanja Schultz.

The authors are with LIMSI-CNRS, BP133, 91403 Orsay cedex, France (e-mail: thomas@l2f-inesc.pt, lamel@limsi.fr, thomas@l2f-inesc.pt, lamel@limsi.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2022295

on the distinctive features specific to the Amharic phonemes. By giving a "phonemic" distance between two lexical units, word splits that result in the largest distances between sub-word units can be favored. The distinctive features are based on very general theoretical sound properties. According to Jakobson, phonemes of a specific language are distinguishable with a small set of articulatory and acoustico–perceptive features, called *distinctive features*, such as voiced-unvoiced property or the place of articulation often corresponding to the point of constriction in the vocal tract [13]. A problem is that splitting words can create homophones or near-homophones, particularly if multiple pronunciation variants are allowed for the lexical units. To overcome this drawback, an additional constraint was introduced to forbid word splits that could have the same pronunciation variant. Results incorporating these properties for vowels were reported in [14], showing an absolute WER reduction of 0.4% relative to the word-based system. However, the experiments in this previous work were carried out on a development corpus, since no additional test corpus was available. In the present article, cross-validation has been used to test the approach. The distinctive feature property has also been extended to the consonants, while in previous work it was limited to the vowels. Finally, complementary experiments with longer-span language models (5-gram) are also reported.

The paper is organized as follows. The next section discusses the key role of words in ASR and motivates the use of sub-word units for MRLs. This is followed by an overview of the ASR literature with sub-word units. Section IV describes the baseline version of the corpus-based word decompounding algorithm Morfessor, with the modifications made to incorporate "oral properties." Section V presents the experimental results carried out on the Amharic corpus. Since morphological decomposition results in the redefinition of words or lexical entries used for ASR, each explored configuration implies renormalization of the available texts and transcripts, as well as the retraining of the language and acoustic models. All modifications in the word decompounding algorithm are fully tested by measuring ASR performance in terms of word error rates in comparison to the reference word-based system. Finally, some conclusions and perspectives are given.

## II. Reference Units for Speech Recognition

Speech recognition consists of finding the best elementary unit sequence $\hat{M}$, which is the hypothesis with the highest probability, given a speech signal $S$: $\hat{M} = \arg\max_{M} P(M|S) = \arg\max_{M} P(S|M)P(M)$. By and large, the most widely used recognition unit is the "word," where the definition of a word may vary across languages and systems. Performance is usually measured by word error rate, which is the sum of all kinds of word errors (insertions $(\#I)$, substitutions $(\#S)$, and deletions $(\#D)$), normalized by the number of words $(N)$ in the reference (manual transcription in general). The WER is formulated as $\mathrm{WER} = (\#I + \#S + \#D)/N$. Word errors are determined by dynamically aligning the recognition hypothesis to the reference transcription at a sentence level. We used the NIST standard scoring tool "sclite," available at http://www.nist.gov/speech/tools.

In Linguistics, the concept of *word* is often described as complex and problematic, with difficulties arising when word identification has to be done.[1] In speech recognition, only words specified in a lexicon can be recognized. So some kind of word segmentation and identification are necessary to build a recognition lexicon, and it is typical to take a very pragmatic approach, identifying words in as simple a manner as possible. Even for languages written with a space or another separator between words, there are normalization choices to make. In French for example, the use of the apostrophe is very frequent, as for the definite article $l'$. Words like $l'oral$ can be considered as two words $l'$ and *oral*, or just one word since there is no space between the two distinct words. In order to avoid increasing substantially the lexicon size, the first possibility may be chosen, and all small words $(c', j', l', m', n', s't', \ldots)$ may be separated from their associated nouns and considered as words. This choice reduces lexical variation at the cost of introducing many words with a single phone. Such normalization issues in French are discussed in [16]. As explained in the Chapter "The use of lexica in Automatic Speech Recognition," by Adda-Decker and Lamel [17], normalization choices for the apostrophe may be different in French and in English. In English, apostrophes are not as frequent as they are in French, and therefore they are typically not considered to be word separators. Contractions like *I'll*, *you've*, or *he's* and as well as compound words and multi-word sequences are often used as lexical entries for speech recognition [18]–[20]. These normalization practices, derived from experience gained by specialists in speech recognition, may be different according to the experts that choose them, but they illustrate well the issues linked with word definition for ASR. The specific choices may also differ depending on the language, task, and application. Dialog systems and conversational speech recognizers have been reported to benefit from using compound words in order to facilitate the use of pronunciation variants specific to conversational speech.

Some languages have no word separators, as it is the case for various Asiatic languages such Chinese, Japanese, and Thai. For these languages, segmentation algorithms are required for pre-processing and/or postprocessing. In general, a reference lexicon is used but very often, multiple word segmentations are possible for the same sentence. Various automatic techniques have been proposed to try to remove this ambiguity, the most popular being the maximum match segmentation, which tries to find the longest words to match the characters in a sentence. In 2005, the "Second International Chinese Word Segmentation Bakeoff" showed that despite performance gains in the word segmentation task, the main issue is still the processing of the OOV words [21]. In order to avoid this issue, the ASR performance is typically measured at character level, with character error rates (CERs) instead of word error rates [22], [23].

For morphologically rich languages, the definition and selection of lexical units is a popular topic in ASR, since prohibitive lexicon sizes would be required to achieve reasonable

[1] Linguists use other concepts, such as *word-form*, *lexeme*, and *autonomous syntagm* [15], for example.

TABLE I
OUT-OF-VOCABULARY RATE (OOV) COMPARISON FOR TWO RICH MORPHOLOGY LANGUAGES (AMHARIC AND TURKISH), AND TWO LANGUAGES THAT HAVE A "LESS RICH" MORPHOLOGY (ENGLISH AND FRENCH)

| *Language* | *Lexicon size (word types)* | *OOV(%)* |
|---|---|---|
| *English* | 65k | 0.6 |
| *French* | 65k | 1.2 |
| *Amharic* | 133k | 6.9 |
| *Turkish* | 250k | 6.5 |

lexical coverage. One characteristic of such languages is the increase in distinct word number ("word types"), as a function of the total number of words of a corpus ("word tokens"). For these languages, the increase is much faster than for other languages. The study reported in [24] for example, distinguishes Finnish, Estonian, Turkish, and Arabic from English on that point. Table I gives lexicon sizes and OOV rates of systems developed at LIMSI for English and French, and for two morphologically rich languages, Amharic and Turkish. Nowadays, it is common practice to use lexicons comprised of at least 65k words and most state-of-the-art recognition system developers consider acceptable OOV rates to be under 1%. As shown in Table I with 65k words the OOV rate for English is 0.6%, and is on the order of 1.2% for French. Using a 200k word lexicon can reduce the OOV rate to under 0.5% for French [25].

For Amharic and Turkish, much higher OOV rates are observed, 6.5% and 6.9% respectively, with substantially larger lexicons. This difference is mainly due to the rich morphology of Amharic and Turkish, but is also accentuated by the lack of resources compared to English and French. In [26], a 96.4 million word text corpus is used to train language models for broadcast news transcription in Turkish. If all observed word forms were included in the lexicon, it would be comprised of 1.4M words, a prohibitive size for speech recognition. Lexicon size reduction is quite interesting in that case, and decompounding words into sub-word units can serve to decrease both the recognition lexicon size and OOV rates. Some illustrations found in the literature are as follows.

- The German word *Schulelternbeiratsmitglieder* was decompounded into *Schuleltern + beiratsmitglieder*, then into *Schul + eltern + beirats + mitglieder*, by using a character-based maximum branching factor algorithm [2].
- The Turkish sentence *Isteklerimizi elde ettik dedi* has been decompounded into *Istekler+ imizi el+ de etti+ k de+ di*, by using the Morfessor algorithm, also used in this work and presented in Section IV[27].

Recently, several sites have reported on morphological decomposition for the Arabic language [28]–[30] where sub-word units such as prefixes (*Al, bAl, fAl, kAl, b, f, k, l, s, w*, . . .) are used to decompound words. Rules are typically applied to restrict the decomposition of frequent words avoiding some possible confusions. These reported experiments were carried out with state-of-the-art systems trained on very large corpora.

Some of the above-mentioned studies showed improvement in recognition performance obtained by word decompounding. The methods used in these studies were different, Section III presents and discusses some of them, along with other studies found in the literature. Here first is a brief introduction to the Amharic language, which is used as a case study in this work.

TABLE II
EXAMPLE OF DIFFERENT ORDERS (SYLLABLE NUCLEUS) ASSOCIATED TO THE 'L' CONSONANT, GIVEN IN AMHARIC SCRIPT AND OUR LATIN TRANSLITERATION. THE '$x$' STANDS FOR A REDUCED VOWEL (SCHWA)

| Ge'ez Symbols | ለ | ሉ | ሊ | ላ | ሌ | ል | ሎ |
|---|---|---|---|---|---|---|---|
| Transliteration | lE | lu | li | la | le | lx | lo |

Amharic was chosen as an example of a Semitic language, language family to which Arabic belongs to. It is mainly spoken in Ethiopia. After Arabic, it is the second most widely spoken Semitic language in the world, with 22 million speakers [31]. Despite its "official working" language status, and its nationwide use, Amharic suffers from poor representation on the Internet, and may be considered as a "less-represented" language, for which only small quantities of written texts are available [32]. For speech recognition, the lack of text resources makes language model probability estimation difficult, and often implies high out-of-vocabulary rates. In Amharic, these problems are increased by its rich and complex morphology, which is inflectional and derivational [33]. One characteristic of languages with a rich morphology is a high increase in the number of word types as a function of the number of word tokens [34]. Reference [9, Table IV] compares the frequencies of word types in Amharic and in English, showing that word type frequencies are quite a bit lower for Amharic.

Amharic has 34 basic symbols, for which there are seven vocalizations (transliterated form): /ɛ/, /u/, /i/, /a/, /e/, /ə/ and /o/,, referred to as the seven orders. The basic symbols are modified in a number of different ways to indicate the different vocalizations. 85% of the syllables represent a CV sequence (C for consonant and V for vowel), one symbol represents the complex sound /ts/V and the remainder represent CwV sequences (where w is a semi-consonant). In this study, Cw has been considered as a single phone. For practical reasons, the Amharic script was transliterated into a set of Latin letters. Table II shows an example of the ለ syllable, that is transliterated by /lE/ corresponding to the phonetic transcription [lϵ], given with its seven orders. The sixth-order syllable nucleus is a schwa, written as "x."

## III. WORD DECOMPOUNDING FOR SPEECH RECOGNITION

The use of sub-word units in speech recognition is not new, with studies dating from the mid 1990s, but it remains an active research area. Most of the studies use "Top-Down" methods: starting from full word forms, words are decompounded into smaller units. Once sub-word units have been selected, the studies differ on how the sub-word units are used in the decoding. Sub-word units can be used at different levels of modeling: acoustic modeling and/or language modeling, for all the decoding, or just during lattice rescoring. Kirchhoff and Sarikaya, who led the ISCA Workshop "Processing Morphologically Rich Languages" during the Interspeech conference in Antwerp, August 2007, distinguished three main approaches that are briefly reviewed here [34].

To better identify the pros and cons of the three methods, they are presented in relation to the generic speech recognition formulation as shown in (1). The likelihood $P(S|M)$ of the signal $S$ given the word sequence $M$, is developed as the sum of the pronunciation probabilities (pronunciation variants named $H$) associated to the word sequence $M$. $H$ corresponds to a sequence of acoustic models (phones in general)

$$\begin{aligned}\hat{M} &= \arg\max_{M} P(S|M)P(M) \\ &= \arg\max_{M} \sum_{H} P(S|H)P(H|M)P(M). \qquad (1)\end{aligned}$$

The first method consists of using sub-word units in all modeling elements of the speech recognizer: acoustic models, lexicon, and language model. This is the approach taken in [3], [35], for example, with application to German and French. If the French word *aller* was decompounded into *all- er*, then with this method, two acoustic models would be used, one for *all-* and one for *er*. The advantage here is the economy of the pronunciation lexicon, since in this case $P(H|M) = P(M|M) = 1$. A limitation of this method is the high complexity of the acoustic models, with a number of states that should depend on the number of phones in each sub-unit.

The second method uses the sub-word units in the lexicon and the language model in the decoding process. The recognition units may be a combination of words and sub-word units. Acoustic units are not based on the sub-word units, but are generally phones or phone-like units. Some studies using this approach are for example [4], [24], [28], [29], [36]. With this method, building a pronunciation lexicon (i.e., determining $H$ for each $M$) is necessary and may pose difficulties for some sub-word units. One solution would be to decompound words into sub-word units for which pronunciations are known or easy to determine. In the literature, languages for which this method has been used have a simple grapheme-to-phoneme conversion. This is globally the case for the Amharic language of interest in this study. Another potential problem source is modeling the $P(S|H)$ term, in the sense that small units are known to increase acoustic–phonetic confusions for the system, and their probabilities $P(S|H)$ are very similar. In the present work, special care in sub-unit generation was taken in order to try to avoid the creation of units that were too small or too similar, with the use of the new "oral" properties, presented in Section IV-C.

The third method uses sub-word units only in a rescoring pass, i.e., a sub-word based language model is used to rescore recognition hypotheses, generally structured as lattices or consensus networks, generated by a word-based system. Examples of this method can be found in [11], [27], [37]. Again concerning (1), only $P(M)$ is modified in comparison to a word-based system. The advantage of this method is double: there is no increase in acoustic–phonetic confusability since words are used during the acoustic part of the decoding, and there is no problem of finding pronunciations for sub-word units. Various strategies can be adopted: simply decompounding words from the N-best hypotheses and rescoring with a sub-unit based language model (LM); combining scores achieved with a word based LM and scores achieved with a sub-unit based LM [11]; or expanding lattices or consensus networks by adding nodes and arcs with words that begin with a same prefix for example [27].

Based on the literature studies, it is not possible to determine which of the three approaches is the best, and the end choice is likely to depend on a variety of factors and constraints. Since the work presented here is for a language that has a straightforward grapheme to phoneme conversion, the second approach, that combines the use of sub-word units for language modeling in all decoding steps with phone-based acoustic units, was chosen. In Section IV, the word decompounding strategy, enhanced for speech recognition purposes, is described.

## IV. Incorporating ASR-Oriented Properties in Corpus-Based Word Decompounding

Automatic word decompounding is investigated as a means to help select recognition units in an almost language-independent manner. In order to minimize the work needed to apply the adopted approach to different languages, a data-driven algorithm, requiring little linguistic knowledge, was explored. Various unsupervised morphology analysis algorithms are open source or easy to implement, such as Harris [38], Goldsmith [39], and Morfessor [7]. The Morfessor algorithm was chosen since it seemed to be a more general model than the others, for example unlike Goldsmith, no assumption about the basic structure of words is made. Furthermore, several recent studies making use of Morfessor reported improvements for a variety of languages, using either the second [4], [24], or the third [27] approach described in the previous section. This work is an extension of the Morfessor algorithm, as implemented in the open source Perl program called "Morfessor 1.0," available at http://www.cis.hut.fi/projects/morpho/.

### A. Baseline Morfessor 1.0 Algorithm

Morfessor is an iterative algorithm that given a corpus, proposes word segmentations found with an optimization criterion. The authors use the term of "morphemes" to name the sub-word units proposed by Morfessor, but they also use the neologism "morphs," since the splits are not always true morphemes in a linguistic sense. Finally, morphs can be either words or word splits.

An overview of the basics of this algorithm is provided here, for further information the reader is referred to [7]. The program has two modes:

1) A "training" mode which creates a word segmentation model given a lexicon with optional frequency counts. Training uses a maximum a posteriori (MAP) criterion based on several text properties, including word frequencies and string probabilities.
2) A "decoding" mode in which a previously learnt decomposition model can be used to decompound a new word list. Each input word is decomposed into a sequence of morphs that exist in the model. This search algorithm maximizes only the morph frequencies, and no retraining is done. Words that are not in the model can be decomposed into a sequence of known morphs.

During model training, the algorithm tries to iteratively maximize the following estimate:

$$M = \arg\max_{L} P(L|\text{corpus}) = \arg\max_{L} \underbrace{P(\text{corpus}|L)}_{\text{Likelihood}} \underbrace{P(L)}_{A\ priori} \quad (2)$$

where $P(\text{corpus}|L)$ is the maximum-likelihood estimate of the corpus given a lexicon $L$, based on the word frequencies, and $P(L)$ is the *a priori* probability of the lexicon $L$, i.e., the probability of getting $M$ distinct morphs $m_1, \ldots, m_M$ as shown in (3). Properties used in the baseline version are morph frequency, morph string, and morph length, respectively, denominated $n(m_k)$, $s(m_1)$, and $l(m_1)$ in (3). For more details about the computation of these terms, the reader is referred to [7]. Our modifications, described in the following sections, affect the *a priori* properties used as

$$P(L) = P(n(m_1), \ldots, n(m_M))P(s(m_1), \ldots, s(m_M))P(l(m_1), \ldots, l(m_M)). \quad (3)$$

As it is common practice for this type of algorithm, probabilities are not multiplied as is, since they are often very small, but the negative log probabilities are summed. Maximizing the likelihood consists then in minimizing a sum of negative log probabilities, which can be seen as minimizing a cost function.

The decoding part of Morfessor is different from the training mode, since chosen morphs are those which minimize a cost function based only on the morph frequencies, and no other property.

In both modes, every word position is a potential candidate for split, and the algorithm explores all word substrings. Words can be split into various morphs, but words are not decompounded if splitting does not reduce the cost function value.

### B. Modified End-of-Word Probability

In the baseline Morfessor program, the character probabilities are static constants, calculated only once during model initialization, as the simple ratio of the number of occurrences of the character divided by the total number of characters in the corpus. These are independent of word position. To represent the word boundary, a space character is added to each lexical entry. The end-of-word probability is the probability of the space character, and has the same value for all words and morphs in the corpus.

Inspired by Harris' algorithm [38] and previous work on German word decomposition [2], we propose replacing this static probability by the probability $P_H$ defined in (4), to take the string context into account. $P(l(m_1), \ldots, l(m_M))$ in (3) is replaced by $P_H(l(m_1), \ldots, l(m_M))$. The word beginning symbol (WB) stands for the strings that begin a given word, from length zero to the word length itself. The probability that a word beginning WB is a morph, is defined as the ratio of the number of distinct letters $L(\text{WB})$ which can follow WB over the total number of distinct letters $L$. The division by $L$ is not mandatory since it is a constant and thus does not influence the cost minimization, but it was kept for coherence, since the other quantities used in the algorithm are probabilities. This term is inspired from Harris' observation that this number decreases naturally from the word start, and that if it increases at a given point in the string, the sub-string up to this point might be a morph, that can be followed by many different suffixes

$$P_H(\text{WB}) = \frac{L(\text{WB})}{L}. \quad (4)$$

This definition favors short morphs, which is potentially interesting for languages where the word compounding generation process corresponds to the addition of prefixes and suffixes that are grammatical morphemes such as pronouns, possessive and demonstrative adjectives, prepositions, and postpositions.

### C. Modified Algorithm for ASR

All the properties used in the Morfessor program are based on written language and do not incorporate any "oral" properties that could be useful for ASR. Two modifications were introduced to try incorporate such properties.

*1) Distinctive Feature Motivated Property:* This property is an attempt to incorporate linguistic knowledge in the decompounding process. A phone-based feature was added to the $P(L)$ term of (2) and (3). This property aims to give an estimation of the phonemic confusability between lexical units. It is theoretical and relies on some distinctive features (DF) of the phones used in the language of study. The DFs are basically the same as those used in the decision tree that merges contexts during acoustic model training (as described in the experimental Section V-D). For a particular morph, the smaller the feature value is, the greater the number of similar morphs (in terms of DFs) there are in the lexicon. As for the other terms of the Morfessor algorithm, it takes the form of a probability.

Equation (5) gives the definition for a morph $m_k$. The DFs of its vowels are compared to the DFs of the vowels of all the other morphs that share the same consonantal root. The compared vowels have the same position in the morphs being compared. The same definition is used for consonants, however in that case, the DFs of morphs that share the same "vocal root" are compared. For example, the two Amharic words with the phoneme transcriptions of [nɛwa], [nɛwə],, share the same $[n, w]$ consonantal root. Thus the vowel DFs are compared. Both words have the same first vowel, which is ignored in the computation, otherwise the feature would be zero. Only the vowel pair [a,ə] will have a contribution. The other possible vowel pairs [ɛ,a] and [ɛ,ə] are not used since they involve vowels that have different word positions. In an analogous manner, if two words share the same "vocal root," then DF differences in the consonants can be computed.

Two distinct results, one for the vowels of morph $m_k$ and one for its consonants can be computed. In the next sections, results using only the vowel DFs, only the consonant DFs, and both DFs (computed by summing their logarithms) will be given.

The following discussion explains how this feature is computed for vowels, the extension to consonants being straightforward. Equations (5) and (6) are used to define how the difference in score of distinctive features for vowels is computed as follows:

$$P_{DF}(m_k) = \prod_{j=1, j \neq k}^{j=N_k} P_{DF}(m_k, m_j) \quad (5)$$

TABLE III
DISTINCTIVE FEATURES OF THE AMHARIC VOWELS, USED WITH THE ALGORITHM. REMARK: BASED ON THE VOWEL CONFUSIONS REPORTED IN A PREVIOUS STUDY [10], FOR THIS STUDY THE /A/ IS CONSIDERED NON-TENSE

| *DF* | *Vowels* | | | | | | |
|---|---|---|---|---|---|---|---|
| *IPA* | ɛ | u | i | a | e | ə | ɔ |
| *high* | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| *low* | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| *round* | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| *tense* | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| *reduced* | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| *back* | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| *long* | 0 | 1 | 1 | 1 | 1 | 0 | 1 |

TABLE IV
DECOMPOSITION OPTIONS COMPARED IN THIS STUDY

| *Option* | *Comment* |
|---|---|
| BL | Baseline word based system, no decompounding |
| M | Baseline Morfessor 1.0 |
| M H | M + modified 'Harris' |
| M H DFV | M H + distinctive features parameter of vowels only |
| M H DFC | M H + distinctive features parameter of consonants only |
| M H DFCV | M H + distinctive features parameter of vowels and consonants |
| Cc | + confusion constraint |

with

$$P_{DF}(m_k, m_j) = \prod_{l=1}^{l=V_k} \frac{\Delta_{kl,jl}}{C} \tag{6}$$

where $N_k$ is the number of morphs that share the same consonantal root, $\Delta_{kl,jl}$ is the number of DFs in which the $l$th vowels of morphs $m_k$ and $m_j$ differ (computed only if the vowels are different), $V_k$ is the total number of vowels in morph $m_k$, and $C$ is the total number of distinct DFs. Note that while $P_{DF}(m_k) \in [0,1]$, $P_{DF}$ is not a probability since it does not sum to one. The more distinct DFs two morphs have, the bigger the feature value is, and the smaller the associated "cost" (negative logarithm of $P_{DF}$) is. This feature thus aims to favor word decompositions that give morphs which have distinct DFs compared to the other morphs.

To evaluate $\Delta_{kl,jl}$, one can use standard DF tables found in phonetics literature, for example in [40]. The distinctive features used in this study concern vowels and consonants, and are given for information in Table III for vowels only. Features for consonants are similar ([41, p. 144]).

Finally, as shown in (7), $P_{DF}$ has been incorporated in $P(L)$ as an additional term. Equation (7) is our modified version of the original $P(L)$ Morfessor formulation, given in (3). As for the other three properties ($n$, $s$, $l$), the property $P_{DF}(m_k)$ is considered to be independent from the other morph feature values so that $P_{DF}(m_1,\ldots,m_M) = \prod_{k=1}^{M} P_{DF}(m_k)$.

$$\begin{aligned} P(L) = P(n(m_1),\ldots,n(m_M))P(s(m_1),\ldots,\\ (m_M))P_H(l(m_1),\ldots,l(m_M))P_{DF}(m_1,\ldots,m_M). \end{aligned} \tag{7}$$

*2) Phonemic Confusion Constraint:* The DF property is theoretical and therefore does not account for the phonological variation observed in real world speech, such as in the choice of vowel alternatives. In [12], syllabotactic alignments were studied in order to determine the most frequent confusions at the syllable level. For each syllable, the vowel that was most often substituted by the aligner was determined. These confusion pairs provide an additional means of reducing phonemic confusion amongst units arising from the decompounding.

During the decompounding process, candidates for word splitting that differ from other morphs by only one syllable are compared. If the pair of syllables is among the most frequently confused pairs found in the alignment study, the candidate is rejected (the split is refused). In the previous example with the two words [nɛwa], [nɛwə],, if the algorithm already split the first word into [n$\epsilon$ + wa], and if the split of the second word into [nɛ+ wə] was found to lower the global function cost and thus be a good decomposition, the Cc constraint would forbid this decomposition if the syllable pair [wa] and [wə] was among the confusion pairs resulting from the syllabotactic alignments.

The different options investigated with the decompounding algorithm are summarized in Table IV. The configurations M, MH, MHDFV, MHDFC, and MHDFCV are compared both with and without the confusion constraint Cc.

## V. EXPERIMENTAL STUDY

In this section, recognition experiments for the Amharic language are reported using a corpus of broadcast news data.

### A. Amharic Corpus

Some recent studies, for example [9], [42], have addressed speech recognition and speech processing for Amharic using read speech. In the experiments reported here, a broadcast news speech corpus is used. Compared to other languages for which models and systems have been developed [43], the available Amharic audio corpus is quite small. It is comprised of 37 h of broadcast news data from two sources, *Deutsche Welle* (25 h 26 min) and *Radio Medhin* (11 h 45 min). The data were transcribed by native Ethiopian speakers, and contain a total of 247k words with 50k distinct lexemes. Two hours of data taken from the latest shows of each source were reserved for development and test. This data contains 14.2 k words, of which almost 15% do not appear in the training portion. In a previous study [14], results were reported on the same 2-h corpus that was used for development purposes. This means that certain parameters, such as the language model interpolation coefficients were optimized on the data potentially introducing a bias. Since no additional data are available, for the experiments reported in this paper, the same 2-h corpus was divided into two distinct subsets, 80% for development, and 20% for test (percentages based on the number of words). Seven distinct dev/test configurations were randomly selected, in order to do a classical cross-validation. Table V gives the number of speakers and words in the different subsets. Depending on the randomly selected files for dev/test, the number of speakers is between 12 and 15 for the dev, 4 and 7 for test.

In addition to the transcriptions of the audio data, about 4.6 million words of newspaper and web texts have been used for language model training. Over 340 k distinct words are found in these texts.

TABLE V
CHARACTERISTICS OF THE AUDIO CORPUS (NUMBER OF HOURS, SPEAKERS, AND TOTAL NUMBER OF WORDS FOR TRAINING, DEV AND TEST)

| | *Training* | *Development* | Test |
|---|---|---|---|
| | 35hr 14min | 1hr 34min | 23min |
| *# Speakers* | 200 | [12-15] | [4-7] |
| *# Words* | 233k | 11.4k | 2.8k |

TABLE VI
NUMBER OF MORPH TYPES IN THE LEXICONS WITH AND WITHOUT "+" FOR DIFFERENT DECOMPOUNDING OPTIONS. (BL: WORD-BASED SYSTEM, M: BASELINE MORFESSOR, H: HARRIS' OPTION, CC: CONFUSION CONSTRAINT, DFV: DISTINCTIVE FEATURES FOR VOWELS, DFC: DISTINCTIVE FEATURES FOR CONSONANTS, DFCV: DISTINCTIVE FEATURES FOR VOWELS AND CONSONANTS)

| *Options* | Lexicon size | |
|---|---|---|
| | with '+' | without '+' |
| BL | 0 | 133384 |
| M | 95937 | 70267 |
| M Cc | 128239 | 109694 |
| M H | 90740 | 65421 |
| M H Cc | 126105 | 107123 |
| M H DFV | 94198 | 69038 |
| M H DFV Cc | 128404 | 110320 |
| M H DFC | 66190 | 50062 |
| M H DFC Cc | 118596 | 101770 |
| M H DFCV | 66250 | 50193 |
| M H DFCV Cc | 107786 | 93573 |

### B. Decompounding the Training Texts

When building a recognition lexicon from training texts, a frequency cutoff is typically applied to get rid of misspelled words and artifacts. In this study the cutoff is applied after decomposition. It should be noted that given the CV structure of the Amharic language, word splits are allowed only after a vowel. First, a decompounding model is built for a reference lexicon, and then this model is used to decompose all words in the corpus without any frequency cutoff. A new reference lexicon is then selected, applying a frequency cutoff: only morphs occurring at least three times are included in the lexicon. The OOV rate may decrease since OOV words may have been decompounded. The number of lexical tokens in the training text corpus is also increased with this method.

An initial 133 k word-based lexicon was selected. It was comprised of the 50 k distinct words in the acoustic training data transcriptions and all words occurring at least three times in the newspaper and web texts. The out-of-vocabulary rate of the development corpus with this word list is almost 7%, which is quite high compared to the OOV rates obtained for well-represented languages which are typically around 1%.

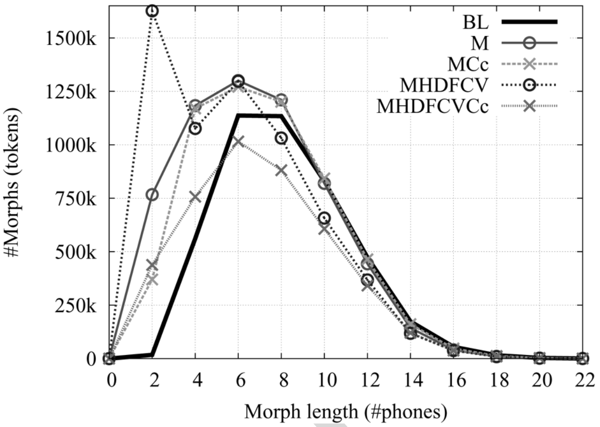

Fig. 1. Number of morph tokens in the training data as a function of the number of phones for different decomposition options. (BL: word-based system, M: baseline Morfessor, Cc: confusion constraint, DFCV: distinctive features for vowels and consonants).

Table VI shows the number of morph types for the different decompounding options listed in Table IV. Since a morph may exist both as a word and as an affix, the explicit use of this information is investigated by adding a "+" sign to prefixes found by the algorithm in order to simplify the work of recombining morphs back into entire words in the ASR experiments. The distinctive feature option for consonants (DFC) gives the smallest lexicon with about 66 k units, being about half the size of the original lexicon. The Cc constraint increases lexicon size by 25%–30% relative to the same configuration without the constraint, except for the DFC option, for which the increase is quite a bit larger (43%). This indicates that the use of DFC splits many words into potentially confusable sub-word units. Since the word and affix entries corresponding to the morph will have the same pronunciations in the recognition lexicon, the choice between forms is made by the language model. The third column gives for information the number of types when no explicit distinction is made between words and affixes (i.e., no "+" sign is added during decomposition). The difference between the second and the third columns is the number of morphs that are also words.

Fig. 1 shows the number of tokens as a function of their length in phones,[2] for different decompounding options. The BL curve (in black) is the baseline curve, with no decompounding. The other curves, for which words were decompounded, show a noticeable shift to smaller word lengths. Some decompounding options have been omitted to keep the figure readable, but these curves are similar to ones shown. The curves with and without the "Cc" option form two distinct groups. As expected, the "non Cc" curves (drawn with "o" points) have substantially more morph tokens with a length of 2 phones compared to the "Cc" curves (drawn with "x" points), since more words are decompounded without the constraint. Basically, the DF property for consonants (DFC) introduces the largest number of small units, and the M H DFCV curve have almost twice as many 2-phone units than the other "non Cc" curves. As was written in the introduction, small units are more error-prone than longer units (see [10], [11]). Reducing their frequency with the phonetic "Cc" constraint is thus promising, but of course results in a larger lexicon size and/or OOV rate.

### C. Language Model and OOV Rates

The language models are Kneser–Ney smoothed four-gram models, and result from the interpolation of two component LMs: one estimated on the web/newspaper texts and the other on the manual transcripts of the audio training data. The interpolation coefficient was optimized for each LM by measuring the

[2] Recall that characters in Amharic correspond to a syllable, so all points are multiples of 2 phones since the lengths are determined from a canonical pronunciation.

TABLE VII
AVERAGED OOV RATES (%) ON THE TEST CORPUS. (BL: WORD-BASED SYSTEM, M: BASELINE MORFESSOR, H: HARRIS' OPTION, Cc: CONFUSION CONSTRAINT, DFV: DISTINCTIVE FEATURES FOR VOWELS, DFC: DISTINCTIVE FEATURES FOR CONSONANTS, DFCV: DISTINCTIVE FEATURES FOR VOWELS AND CONSONANTS)

| *Options* | *OOV Tokens (%)* |
|---|---|
| BL | 6.83 |
| M | 3.99 |
| M Cc | 4.32 |
| M H | 3.97 |
| M H Cc | 4.30 |
| M H DFV | 3.97 |
| M H DFV Cc | 4.35 |
| M H DFC | 3.47 |
| M H DFC Cc | 4.10 |
| M H DFCV | 3.47 |
| M H DFCV Cc | 4.35 |

perplexity on the development transcripts. Different LMs were built for each set of decompounding options, and for each development/test subdivision. Since some of the words which are not in the baseline vocabulary are decomposed, the OOV rates are reduced. Table VII gives the mean token OOV rates averaged across the seven different test subsets (each about 2.8 k words). The relative reduction in OOV rate ranges from 35% to 50% depending on the options.

### *D. ASR Experiments*

This section reports recognition results obtained with systems trained for each of the decomposition option configurations. The baseline system is the word-based system. The speech recognizers all have two decoding passes, with unsupervised acoustic model adaptation (MLLR) after the first decoding pass [44]. The acoustic models are all tied-state triphone HMMs, covering both word-internal and cross-word contexts, with three states per model and 32 Gaussians per state. State tying is based on classical decision tree clustering, with backoff on diphones and monophones. The set of questions concern the phone position, the distinctive features (and identities) of the phone and the neighboring phones. Since different decompositions result in different recognition units (and therefore different word positions), it was necessary to build specific acoustic models for each set of options. In all cases both intra- and cross-recognition unit contexts are modeled. All acoustic model sets cover about 10.5 k distinct contexts, with a total of about 8.5 k tied states.

Table VIII gives the OOV and word error rates (WER) for the different ASR systems, for the seven development/test configurations, estimated after recombining prefixes (that end with a "+" sign) and roots back into full words. The means of the WERs over the seven configurations are given in the last column. The OOV rate for the word-based system ranges from 5.6% to 8.7%, with an average of 6.8%, which is close to that of the full development data set (6.9%) used in [14]. The largest OOV rate is for subset 3, and the smallest rates are for subsets 2 and 5. The full-word baseline system has a mean WER of 23.6%. The five systems M, MH and MHDFV, MHDFC, MHDFCV, which do not use the confusion constraint Cc, perform slightly less well than the baseline system. On the contrary, the five Cc systems all give small gains. The confusion constraints between lexical units appears to be useful for identifying recognition units when used in conjunction with word decompounding. The worst performance is obtained by the MHDFCV system, which is the algorithm that splits the largest number of words. This result illustrates well the compromise between OOV rate reduction and increased confusability between lexical units when decompounding is used.

TABLE VIII
WORD ERROR RATES FOR THE DIFFERENT ASR SYSTEMS. (BL: WORD-BASED SYSTEM, M: BASELINE MORFESSOR, H: HARRIS' OPTION, Cc: CONFUSION CONSTRAINT, DFV: DISTINCTIVE FEATURES FOR VOWELS, DFC: DISTINCTIVE FEATURES FOR CONSONANTS, DFCV: DISTINCTIVE FEATURES FOR VOWELS AND CONSONANTS). OOV RATES WITH THE INITIAL 133 K LEXICON ARE ALSO GIVEN FOR EACH BATCH

|  | *WER (%)* |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| *Algorithm* | *Batch* |  |  |  |  |  |  |  |
| *Options* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | *Mean* |
| OOV | 7.3 | 5.6 | 8.7 | 6.7 | 5.6 | 6.7 | 7.4 | 6.8 |
| BL | 23.8 | 19.8 | 24.1 | 24.3 | 23.2 | 22.3 | 27.4 | 23.6 |
| M | 24.8 | 20.1 | 23.4 | 24.8 | 23.4 | 22.6 | 28.2 | 23.9 |
| M Cc | 23.4 | 19.6 | 22.7 | 23.8 | 22.5 | 21.3 | 27.5 | 23.0 |
| M H | 24.9 | 20.2 | 23.3 | 24.7 | 23.5 | 23.0 | 28.5 | 24.0 |
| M H Cc | 23.9 | 19.3 | 23.2 | 23.8 | 22.3 | 21.3 | 27.0 | 23.0 |
| M H DFV | 24.6 | 20.2 | 22.9 | 24.6 | 23.4 | 22.4 | 28.8 | 23.8 |
| M H DFV Cc | 23.5 | 19.6 | 23.0 | 23.6 | 22.5 | 21.5 | 27.2 | 22.9 |
| M H DFC | 24.9 | 20.0 | 23.5 | 25.0 | 23.9 | 22.2 | 27.5 | 23.8 |
| M H DFC Cc | 24.0 | 19.5 | 22.7 | 23.7 | 22.5 | 21.0 | 27.2 | 22.9 |
| M H DFCV | 24.8 | 20.3 | 23.4 | 25.3 | 24.0 | 22.2 | 28.9 | 24.1 |
| M H DFCV Cc | 23.6 | 19.9 | 22.9 | 24.3 | 22.9 | 21.3 | 27.0 | 23.1 |

The Harris modification seems useful since it produces smaller lexicons than with Morfessor baseline, and the same mean WER is obtained when using the Cc option (23% WER for both MCC and MHCc systems). Concerning the DF option, there is a 0.4% absolute WER reduction between the MHDFV, MHDFC and MHDFCV systems and their corresponding Cc version. The best performance is obtained with the DFV and DFC motivated systems (MHDFVCc and MHDFCCc) which achieves a 0.7% absolute improvement compared to the baseline. Nevertheless, the WERs of the two systems vary depending on the dev/test subdivision, which can surely be attributed to the small size of the individual sets. It can be seen that results on batch number 3 are different from the other batches in that all the morph-based systems performed better than the word-based system. This may be due to the higher OOV rate of this subset (8.7%) with the baseline system. Significance tests at word-level (MAPSSWE) have been conducted with the "sc_stats" NIST tool, available at www.nist.gov/speech/tools. In comparison with the word-based system, the system based on the baseline Morfessor algorithm does not show any significative difference for any of the batches, although it performs slightly worse on all the test sets, with the exception of batch 3. The two best systems (MHDFVCc and MHDFCCc) show significative differences in performance with the classical 95% confidence threshold, only for batch number 3. For test sets 4, 5, and 6, the threshold is about 85%, and for the others, the performance difference is not significative. This indicates that the modifications seem more useful with test sets that present the highest OOV rates.

By comparing the distinct types of errors, with the percentages of insertions, deletions and substitutions, it appears that

TABLE IX
COMPARISON OF THE WORD ERROR RATES (WER) FOR THE MHDFV AND MHDFVCC SYSTEMS WITH FINAL 5-GRAM LM RESCORING WITH THE WER OF THE CORRESPONDING PREVIOUSLY USED SYSTEMS (4-GRAM LM). OOV RATES WITH WORD-BASED SYSTEMS AND PREVIOUS WERS OBTAINED WITH 4-GRAM RESCORING ARE ALSO GIVEN. IMPROVEMENTS (WER REDUCTIONS) ARE SHOWN IN BLUE AND DEGRADATIONS (WER INCREASES) ARE SHOWN IN RED (M: BASELINE MORFESSOR, H: HARRIS' OPTION, CC: CONFUSION CONSTRAINT, DFV: DISTINCTIVE FEATURES FOR VOWELS)

| | *WER(%)* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Algorithm* | *Batch* | | | | | | | |
| *Options* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | *Mean* |
| OOV (%) | 7.3 | 5.6 | 8.7 | 6.7 | 5.6 | 6.7 | 7.4 | 6.8 |
| M H DFV 4g | 24.6 | 20.2 | 22.9 | 24.6 | 23.4 | 22.4 | 28.8 | 23.8 |
| M H DFV 5g | 24.3 / -0.3 | 20.3 / +0.1 | 22.9 / 0.0 | 24.4 / -0.2 | 23.3 / -0.1 | 22.3 / -0.1 | 28.5 / -0.3 | 23.7 / -0.13 |
| M H DFV Cc 4g | 23.5 | 19.6 | 23.0 | 23.6 | 22.5 | 21.5 | 27.2 | 22.9 |
| M H DFV Cc 5g | 23.7 / +0.2 | 19.4 / -0.2 | 23.0 / 0.0 | 23.7 / +0.1 | 22.4 / -0.1 | 21.3 / -0.2 | 27.1 / -0.1 | 22.9 / -0.04 |

all the morph-based systems have higher average deletion rates than the baseline (2.2% for the BL system versus 2.8% for the M system for example), but lower insertion rates (2.4% for the BL system, 2.1% for the M system). Systems which do not use the Cc constraint have higher substitution rates, suggesting that the Cc constraint is doing what it was designed to do. Looking at the decoder output, the systems (without Cc) do have a tendency to insert small morphs. However this effect is lost after recombining morphs into words. When the morphs are glued together, the errors are counted as substitutions when compared to the word based reference.

The two best systems (MHDFVCc and MHDFCCc) have similar insertion plus deletion rates as the baseline, but the substitution rate is a bit smaller (18.3% vs 19.0%). This improvement may be explained by the recognition of ex-OOV words, as analyzed in the next paragraph.

It was mentioned earlier that word decompounding possibly allows words that were OOV before decompounding to be recognized since sub-word units can be combined to form a word that was not in the initial lexicon. Using batch number 1 for analysis, the initial test OOV rate is 7.3% with 242 OOV tokens for a total of 3321 words. For all the sub-word based systems, about 80 of the OOV words are covered by the respective lexicon. Depending upon the system configuration 26 to 30 of these words were correctly recognized. For batch number 3, which has the highest OOV rate (8.7%), the number of words that are no longer OOV is larger (between 89 to 104 words depending on the system options). More than a half of these words were correctly recognized. For example, with the MH system, 55 ex-OOV words are correctly recognized. There are 2708 words in the associated reference transcripts for this batch, which would suggest that an absolute gain of 2.0% should have been observed. However, as can be seen in Table VIII, the gain is lower, only 0.8% absolute, therefore new errors, i.e., some that were not produced by the word-based system, are introduced by the use of the sub-word units, increasing the error rate by 1.2% absolute. Additional errors may be due to ungrammatical morph sequences, corresponding to the phenomenon called "over-generation." For batch number one for example, the M system output 116 words that are not in the baseline word-based lexicon. 27 of these words were correctly recognized, and correspond to some ex-OOV words as explained above. The remaining 89 words are possibly the result of over-generation, allowing an upper limit on the errors due to over-generation to be estimated at 2.7%. For the MCc system, that creates less decompositions, this estimated upper limit is lower (1.9%). Looking at some of these words, it is clear that these values overestimate the number of introduced errors, since the great majority of these words were already misrecognized with the baseline system. Finally, considering the very permissive rules of Amharic orthography, only an Amharic expert can identify ungrammatical morph sequences properly [45].

When sub-word units are used, the effective span of an n-gram language model is reduced. Shorter units naturally require longer n-grams. In [4] for example, speech recognition experiments were carried out with 5-gram, 7-gram, and 8-gram LMs for respectively Turkish, Finnish and Estonian. The results reported previously in this section all used a 4-gram span for the language models, which may favor the word-based system. A complementary experiment with 5-gram LMs has been conducted.

In this experiment, a rescoring with 5-gram LMs of the best hypotheses is carried out, using the lattices generated by the previously used decoders (these lattices were generated by 4-gram LMs). Results are reported for the MHDFV Cc system, that had the best recognition performance, and with the MHDFV system to evaluate the impact of the Cc option. Table IX gives the absolute differences between WERs obtained with 5-gram rescoring, in comparison to the previous WERs obtained with 4-gram rescoring, reminded in the table. The lattices used for rescoring are identical, the only change is in the order of the LMs. In the table, improvements are shown in blue and increases in WER are shown in red. Globally, for both options, there are more improvements than degradations in performance, but the differences are quite small. The average improvement for the MHDFV system is larger than for the corresponding system with the Cc option (0.13% against 0.04% in mean), which seems natural since there are more word decompositions without the Cc constraint, and therefore more small morphs that can benefit from a longer-span LM. However for both systems, the differences are very small, and further experiments with longer-span LMs would be needed to draw firm conclusions.

## VI. DISCUSSION

In this paper, sub-word units have been investigated to address the issue of very large lexical variety found in morphologically-rich languages, for the task of automatic speech recognition of broadcast news data. An unsupervised data-driven word decompounding algorithm, which extends the Morfessor algorithm to better suit speech recognition, has been described. The original and modified algorithms have been tested in recognition experiments, where OOV and WER reductions have been obtained on a morphologically rich and less-represented language in which grammatical morphemes are glued to roots. For some systems, gains in performance were achieved since words that were out-of-vocabulary with respect to an initial word lexicon were able to be recognized by the morphologically decompounded ones. Nevertheless, the use of small units, often referred to as "morphs," introduces new errors due to an increased confusability between lexical units. This article has attempted to address both the problem of high OOV rates observed for morphologically rich languages, and the problem of increased confusability when using sub-word units as recognition units.

The "Morfessor" algorithm splits words into smaller units in an iterative manner by maximizing a MAP estimate of a lexicon given a word list with frequency counts. The end-of-word probability computation has been modified to allow more splits. A new phonemic-based parameter motivated by distinctive features (DF), was incorporated as well as phonemic confusion constraints derived from previous experiments with automatic alignment of audio data. Systems built with different combinations of options were compared. The best systems all include the confusion constraint, and the phonemic-based DF parameter for consonants or for vowels. For these systems, the lexicon sizes are reduced by about 10%, along with a small absolute gain in WER (0.7%) relative to the reference error (23.6%) of the word-based system. Without the confusion constraints, all systems had slightly worse performance than the baseline. This result demonstrates the usefulness of the confusion constraints. Without the constraint, small units (2 phones long) are very frequent and somewhat error-prone. The differences in errors of the word based and sub-word based systems were analyzed in order to assess how successful the approach is recovering errors due to words that were OOV for the word-based system. The recognition of "ex-OOV" words gives around a 2% absolute gain, but since new errors are introduced, the overall gain is smaller. Contrastive experiments with longer-span LMs (5-gram LMs) were conducted, but showed very little improvement over the 4-gram LMs used throughout this work.

The DF parameter is a phonetically motivated parameter, introduced for vowels and for consonants. Further investigation should be carried out to confirm the usefulness of this parameter. In the current implementation, the different terms in the MAP estimate are summed, however it may be useful to weight these terms in order to optimize each contribution. Future plans are to test the algorithm on another language similar to Amharic, Arabic for instance, for which ample training data are available, as well as on a language in which the word compounding generation process is even more important, such as German or Turkish.

## References

[1] J.-L. Gauvain, L. Lamel, G. Adda, and M. Adda-Decker, "Speaker-independent continuous speech dictation," in *Speech Commun.*, 1994, vol. 15, pp. 21–37.

[2] M. Adda-Decker, "A corpus-based decompounding algorithm for German lexical modeling in LVCSR," in *Proc. Eurospeech*, Geneva, 2003, pp. 257–260.

[3] P. Geutner, "Using morphology towards better large-vocabulary speech recognition systems," in *Proc. ICASSP*, Detroit, 1995, pp. 445–448.

[4] ***[AUTHOR: Please provide page range]***M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and M. Saraclar, "Unlimited vocabulary speech recognition for agglutinative languages," in *Proc. HLT-NAACL*, New York, 2006.

[5] ***[AUTHOR: Please provide page range]***V.-B. Le, L. Besacier, S. Seng, B. Bigi, and T.-N.-D. Do, "Recent advances in automatic speech recognition for vietnamese," in *Proc. SLTU*, Hanoi, Vietman, 2008.

[6] R. Ordelman, A. van Hessen, and F. de Jong, "Compound decomposition in Dutch large vocabulary speech recognition," in *Proc. Eurospeech*, Geneva, 2003, pp. 225–228.

[7] M. Creutz and K. Lagus, "Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0," Computer and Information Science, Report A81, 2005.

[8] T. Schultz, A. Black, S. Badaskar, M. Hornyak, and J. Kominek, "SPICE: Web-based tools for rapid language adaptation in speech processing systems," in *Proc. Interspeech*, Antwerp, 2007, pp. 2125–2128.

[9] S. Abate and W. Menzel, "Automatic speech recognition for an under-resourced language—Amharic," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 1541–1544.

[10] T. Pellegrini and L. Lamel, "Investigating automatic decomposition for ASR in less represented languages," in *Proc. Interspeech*, Pittsburgh, PA, 2006, pp. 285–288.

[11] K. Kirchhoff, J. Bilmes, J. Henderson, and R. Schwartz, "Novel speech recognition models for Arabic," in *Johns Hopkins University Summer Research Workshop*, 2002, Final report.

[12] T. Pellegrini and L. Lamel, "Experimental detection of vowel pronunciation variants in Amharic," in *Proc. LREC*, Genoa, Italy, 2006, pp. 1005–1008.

[13] R. Jakobson, G. Fant, and M. Halle, *Preliminaries to Speech Analysis*. Cambridge, MA: MIT Press, 1952.

[14] T. Pellegrini and L. Lamel, "Using phonetic features in unsupervised word decompounding for asr with application to a less-represented language," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 1797–1800.

[15] A. Martinet, *Éléments de Linguistique Générale*. Paris, France: Armand Colin, 1980.

[16] G. Adda, M. Adda-Decker, J.-L. Gauvain, and L. Lamel, "Text normalization and speech recognition in French," in *Proc. EuroSpeech*, Rhodes, Greece, 1997, vol. 5, pp. 2711–2714.

[17] F. Van Eynde and D. Gibbon, *Lexicon Development for Speech and Language Processing*. Dordrecht, The Netherlands: Kluwer, 2000.

[18] M. Finke and A. Waibel, "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition," in *Proc. Eurospeech*, Rhodes, 1997, pp. 1963–1966.

[19] J.-L. Gauvain, G. Adda, L. Lamel, and M. Adda-Decker, "Transcribing broadcast news: The Limsi Nov96 Hub4 System," in *Proc. ARPA*, Chantilly, France, 1997, pp. 56–63.

[20] ***[AUTHOR: Please provide page range]***A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. G. Rao, M. Plauche, C. Richey, E. Shriberg, K. Sonmez, F. Weng, and J. Zheng, "The SRI March 2000 hub-5 conversational speech transcription system," in *Proc. NIST Speech Transcription Workshop*, College Park, MD, 2000.

[21] ***[AUTHOR: Please provide page range]***T. Emerson, "The second international chinese word segmentation bakeoff," in *Proc. 4th SIGHAN Workshop Chinese Lang. Process.*, Jeju, Korea, 2005.

[22] L. Chen, L. Lamel, and J.-L. Gauvain, "Transcribing mandarin broadcast news," in *Proc. IEEE Workshop Autom. Speech Recognition*, St. Thomas, Virgin Islands, 2003, pp. 99–104.

[23] M.-Y. Hwang, X. Lei, W. Wang, and T. Shinozaki, "Investigation on Mandarin broadcast news speech recognition," in *Proc. ICSLP*, Pittsburgh, PA, 2006, pp. 1233–1236.

[24] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraclar, and A. Stolcke, "Morph-based speech recognition and modeling of out-of-vocabulary words across languages," *ACM Trans. Speech Lang. Process.*, vol. 5.1 Article 3, 2007.

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

[26] E. Arisoy, H. Sak, and M. Saraclar, "Language modeling for automatic Turkish broadcast news transcription," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 2381–2384.

[27] E. Arisoy and M. Saraclar, "Lattice extension and vocabulary adaptation for Turkish LVCSR," *Speech Lang. Process.*, vol. 10, no. 1, 2009.

[28] M. Afify, R. Sarikaya, H.-K. Kuo, L. Besacier, and Y. Gao, "On the use of morphological analysis for dialectal Arabic speech recognition," in *Proc. ICSLP*, Pittsburgh, PA, 2006, pp. 277–280.

[29] B. Xiang, K. Nguyen, L. Nguyen, R. Schwartz, and J. Makhoul, "Morphological decomposition for Arabic broadcast news transcription," in *Proc. ICASSP*, Toulouse, France, 2006, vol. I, pp. 1089–1092.

[30] L. Lamel, A. Messaoudi, and J.-L. Gauvain, "Investigating morphological decomposition for transcription of Arabic broadcast news and broadcast conversation data," in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 1429–1432.

[31] Omniglot, "A guide to the languages, alphabets, syllabaries and other writing systems of the world." [Online]. Available: http://www.omniglot.com/

[32] L. Asker, A. Argaw, B. Gambäck, and M. Sahlgren, "Applying machine learning to Amharic text classification," in *Proc. 5th World Congr. African Linguist.*, Cologne, Germany, 2007.

[33] D. Appleyard, *Colloquial Amharic*. London, U.K.: Routledge, 1995.

[34] ***[AUTHOR: Please provide page range]***K. Kirchhoff and R. Sarikaya, "Processing morphologically rich languages," in *Proc. Workshop Interspeech*, Antwerp, Belgium, 2007.

[35] P. Geutner, M. Finke, and A. Waibel, "Phonetic-distance-based hypothesis driven lexical adaptation for transcribing multilingual broadcast news," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 771–774.

[36] P. Geutner, C. Carki, and T. Schultz, "Turkish LVCSR: Towards better speech recognition for agglutinative languages," in *Proc. ICASSP*, Istanbul, Turkey, 2000, pp. 1563–1566.

[37] E. Arisoy, H. Dutagaci, and L. Arslan, "A unified language model for large vocabulary continuous speech recognition of Turkish," *Signal Process.*, vol. 86, no. 10, pp. 2844–2862, 2006.

[38] Z. Harris, "From phoneme to morpheme," in *Language*, 1955, vol. 31, pp. 190–222.

[39] J. Goldsmith, "Unsupervised learning of the morphology of a natural language," *Comput. linguist.*, vol. 27, no. 2, pp. 153–198, 2001.

[40] M. Halle and G. Clements, *Problem Book in Phonology*. Cambridge, MA: MIT Press, 1983.

[41] T. Pellegrini, "Transcription Automatique de Langues peu Dotées" Ph.D., Paris-Sud Univ., Orsay, 2008.

[42] S. Abate, W. Menzel, and B. Tafila, "An Amharic speech corpus for large vocabulary continuous speech recognition," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 1601–1604.

[43] L. Lamel, J.-L. Gauvain, G. Adda, M. Adda-Decker, L. Canseco, L. Chen, O. Galibert, A. Messaoudi, and H. Schwenk, "Speech transcription in multiple languages," in *Proc. ICASSP*, Montreal, QC, Canada, 2004, vol. 3, pp. 757–760.

[44] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Commun.*, vol. 37, no. 1–2, pp. 89–108, 2002.

[45] ***[AUTHOR: Please provide page range]***D. Yacob, "Application of the double metaphone algorithm to Amharic orthography," in *Proc. Int. Conf. Ethiopian Studies XV*, Cologne, Germany, 2003.

**Thomas Pellegrini** graduated with a degree in physics from the Ecole Supérieure de Physique et de Chimie Industrielles (ESPCI), Paris, France, in 2003 and received the M.Sc. degree in acoustics, signal processing, computer science applied to music from IRCAM, Paris VI University (ATIAM), in 2003 and the Ph.D. degree in computer science on speech recognition for less-represented languages from LIMSI-CNRS from the Paris-Sud University, in 2008.

He was a Teaching Assistant at Paris La Sorbonne from 2007–2008, teaching mainly object-oriented programming. Since September 2008, he has been a Postdoctoral Researcher at the Spoken Language Systems Lab (L$^2$F), Lisbon, Portugal. His research interests are automatic speech recognition, audio, music, and sound in general.

**Lori Lamel** (M'88) received the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, in 1988.

She is a Senior CNRS Researcher in the Spoken Language Processing group, LIMSI, Orsay, France, which she joined in October 1991. Her principal research activities are in speech recognition, studies in acoustic–phonetics, lexical and phonological modeling, and speaker and language identification. She has been a prime contributor to the LIMSI participations in DARPA benchmark evaluations and developed the American English pronunciation lexicon. She has been involved in many European projects, most recently the IPs Chil, and TCStar. She has over 200 reviewed publications.

Dr. Lamel is a member of the Speech Communication Editorial Board and the Interspeech International Advisory Council. She was a member of the IEEE Signal Processing Society's Speech Technical Committee from 1994 to 1998, and the Advisory Committee of the AFCP, the IEEE James L. Flanagan Speech and Audio Processing Award Committee (2006–2009) and the EU-NSF Working Group for "Spoken-Word Digital Audio Collections." She is a corecipient of the 2004 ISCA Best Paper Award for a paper in the *Speech Communication Journal*.