SPEECH ALIGNMENT AND RECOGNITION EXPERIMENTS FOR LUXEMBOURGISH

Martine Adda-Decker[1,2] *Lori Lamel*[2] *Gilles Adda*[2]

(1) LPP, CNRS-Paris 3/Sorbonne Nouvelle
(2) Groupe TLP, LIMSI-CNRS

ABSTRACT

Luxembourgish, embedded in a multilingual context on the divide between Romance and Germanic cultures, remains one of Europe's under-described languages. In this paper, we propose to study acoustic similarities between Luxembourgish and major contact languages (German, French, English) with the help of automatic speech alignment and recognition systems. Experiments were run using monolingual acoustic models trained on German, French and English together with (i) "multilingual" models trained on pooled speech data from these three languages, or with (ii) native Luxembourgish acoustic models from 1200 hours of untranscribed Luxembourgish audio data using unsupervised methods. We investigated whether Luxembourgish was globally better represented by one of the individual languages, by the multilingual model or by the native (unsupervised) model. While German provides globally the best acoustic match for native Luxembourgish, detailed analyses reveal language-specific preferences, in particular English and Luxembourgish models are preferred on diphthongs. The first ASR results illustrate the accuracy of the various sets of supervised monolingual and multilingual models versus unsupervised Luxembourgish acoustic models. The ASR word error rate is progressively reduced from 60 to 25% on the development data set by unsupervised training of larger context-dependent models on increasing anounts of audio data.

Index Terms— under-resourced languages, languages in contact, Luxembourgish, language similarity, acoustic modeling, multilingual models, large vocabulary speech recognition, forced alignment, unsupervised training.

1. INTRODUCTION

Luxembourg, a small country of less than 500,000 inhabitants in the center of Western Europe, is composed of about 60% of native inhabitants and 40% of immigrants. The national language, Luxembourgish ("Lëtzebuergesch"), has only been considered as an official language since 1984 and is spoken by natives [1]. The immigrant population generally speaks one of Luxembourg's other official languages: French or German. Recently, English has joined the set of prestigious languages of communication, mainly in professional environments.

As pointed out by [2] and [3], Luxembourgish may be considered as a partially under-resourced language, due to the fact that the written production in Luxembourgish remains low even though progress can be observed in recent years. In particular, beyond an increasing literary production in Luxembourgish, new communication media like internet (blogs, social media, news, comments, emails) and smartphones (SMS) foster written production in native Luxembourgish. We may notice an increase in production of Luxembourgish news and commentaries (e.g. www.rtl.lu) and linguistic knowledge and resources, such as lexica and grammars, however, their overall status is certainly lower than for major Western European languages. Although Unesco has classified Luxembourgish as a vulnerable language in 2009, other sources tend to be more optimistic about the development and liveliness of Luxembourgish: the number of speakers (as first and second language) seems to be increasing as well as the volume of written production, even though this production is poorly normalised. As a matter of fact, written Luxembourgish is not systematically taught to children in primary school: German is usually the first written language learned, followed by French. Moreover, writing conventions are extremely liberal and respectful of pronunciation variation.

This paper is a follow-up study of our earlier work which aims at taking Luxembourgish on board as an e-language: an electronically searchable spoken language. In support of this goal, Luxembourgish was added to the list of the European languages of the QUAERO[1] project with the agenda of developing and evaluating an automatic speech transcription system before the end of the project (end of 2013). We report on part of this work here, putting our focus on acoustic modelling of Luxembourgish. We compare different sets of acoustic models: seed models derived from major Western European languages in contact with Luxembourgish, namely German, French and English and native Luxembourgish models estimated via an unsupervised training process [4]. The main research issues we aim to address here are the following:

1. how do the unsupervised native Luxembourgish models compare to the supervised monolingual imported

[1] www.quaero.org

(German, French, English) models and the multilingual models (from pooled audio data of the three languages) in a multiple choice forced alignment setup?

2. how do the unsupervised native Luxembourgish models perform in automatic speech transcription?

With respect to question (1), we expect that for some classes of consonants, such as unvoiced plosives (/p/, /t/ and /k/) which occur both in Luxembourgish and the three considered contact languages, the supervised German, French and English models cover well the Luxembourgish productions and might be preferred to the unsupervised Luxembourgish models during forced alignment, and this even more so as unsupervised training introduces some labelling and segmentation noise to the training data. However, Luxembourgish specific phonemes (e.g. /ɪə/ and /ʊə/) which are not in the other languages' inventories might be better represented by the native models.

Question (2) relates to a more general question of achieved acoustic model accuracy when using an unsupervised training framework. Will we be able to achieve low word error rates, where by low we mean comparable to other European languages on similar data sets. The issue of Luxembourgish may be considered as particularly challenging as the degree of standardisation of both spoken and written language can be considered as rather weak and as Luxembourgish is influenced by the practised contact languages.

These issues have important implications for acoustic modeling in automatic speech recognition and in handling pronunciation variants. In a longer term view, the reported work may also have implications on foreign language teaching and learning, as the proposed multilingual forced alignment gives estimates of acoustic distances between similar phonemes of different languages.

The next section provides a rapid overview of the phonemic inventory of Luxembourgish and its correspondance with the three considered contact languages (German, French, and English). Section 3 presents the different types of acoustic models used in the alignment experiments. Section 4 gives an overview of the multilingual forced alignment process and corresponding results are shown in Section 5. The second research issue is addressed in Section 6. Finally, Section 7 summarizes the results and discusses future challenges for speech technology and linguistic studies of Luxembourgish.

2. PHONEMIC INVENTORY OF LUXEMBOURGISH

The adopted Luxembourgish phonemic inventory includes a total of 60 phonemic symbols including 3 extra-phonemic symbols (for silence, breath and hesitations). Table 1 presents a selection of the phonemic inventory together with illustrating examples (see [1, 5] for more information on the phonemic inventory of Luxembourgish). Luxembourgish is characterized by a particularly high number of diphthongs. To minimize the phonemic inventory size, we could have chosen to code diphthongs using two consecutive symbols, one for the nucleus and one for the offglide (e.g. the sequence /ɑ/ and /j/ for diphthong ɑ͡ɪ). We preferred, however, the option of coding diphthongs and affricates using specific unique symbols. Given the importance of French imports, nasal vowels were included in the inventory, although they are not required for typical Luxembourgish words. Furthermore, native Luxembourgish makes use of a rather complex set of voiced/unvoiced fricatives.

Carrier word (Eng)	Lux	Fre	Ger	Eng
ORAL VOWELS				
l**ii**cht (light)	i	i	i	i
l**i**ddereg (lazy)	ɪ	i	ɪ	ɪ
Leed (suffering)	e:	e	e:	e
DIPHTHONGS				
l**éi**en (to tell lies)	e͡ɪ	e	e	e
l**ou**nen (to hire)	ɔ͡ʊ	o	o	o
l**ie**wen (to live)	ɪ͡ə	i	ɪ	ɪ
l**ue**wen (to praise)	ʊ͡ə	u	ʊ	ʊ
CONSONANTS				
Pad (path)	p	p	p	p
Schoul (school)	ʃ	ʃ	ʃ	ʃ
Gilet (waistcoat)	ʒ	ʒ	ʒ	ʒ

Table 1. Sample cross-lingual phoneme mappings: Lux. targets mapped to same or similar phonemes in 3 contact languages (Fre, Ger, Eng).

3. ACOUSTIC MODELS FOR LUXEMBOURGISH

The need to develop acoustic models for under-resourced languages has already been addressed in previous research [6, 7]. In the current study, we use several types of acoustic models trained either from audio data from one the three contact languages, from the pooled audio data of these three languages, or from Luxembourgish audio data. Whereas the models from the contact languages' data were trained using manual transcriptions (supervised mode), the native Luxembourgish data were trained from automatically transcribed data. For more details on the unsupervised Luxembourgish models, please refer to Section 6.

In this section, we provide some more details about the three sets of context-independent acoustic models borrowed from each of the three well-resourced contact language (German, French , English) as described in [7]. The models were trained on manually transcribed audio data (between 40 and 150 hours) from a variety of audio sources, using language-specific phone sets. The amount of data used to train the native acoustic models and the number of phonemes per language are given in Table 2 (left). Each phone model

Language	*#native phon.*	*training data (h)*	*#Pseudo-Lux. ac. models*
English	48	150	60
French	37	150	60
German	49	40	60
Pooled	-	340	60

Table 2. Phoneme and training data information (in hours) for (supervised) monolingual acoustic models from English, French, German and the multilingual (from pooled audio data) acoustic models.

is a tied-state left-to-right, 3-state CDHMM with Gaussian mixture observation densities (typically containing 64 components). The three sets of pseudo-Luxembourgish acoustic models, each including 60 phones correspond to the English, French and German models and mapping the Luxembourgish phonemes to a close equivalent in each of the contact languages' model sets. Table 1 shows a sample of the adopted cross-lingual mappings that were used to initialize the monolingual models for Luxembourgish. Some symbols are used several times for different Luxembourgish phonemes. For Luxembourgish diphthongs that are missing in the other languages, phonemes corresponding to the nucleus vowel were chosen. An additional set of multiligual acoustic models was trained using the pooled German, French and English audio data that were labeled using their respective IPA correspondances (as exemplified in Table 1).

4. MULTILINGUAL FORCED ALIGNMENT

A forced alignment process was used to investigate whether Luxembourgish was globally better represented by one of the individual contact languages (German, French or English), by the multilingual model stemming from the pooled audio data or by the native (unsupervised) Luxembourgish model. The forced alignment process was adapted to enable each spoken word and even each sound to be aligned with the acoustic model of the best matching language.

Figure 1 shows the implementation of this process via multiple pronunciations in the lexicon. Each phonemic position in a pronunciation may be modeled via a corresponding acoustic model of any type of the available acoustic model sets. The forced alignment consists in selecting the best matching path across the trellis, and each state of this path will indicate the best matching model type (English, French, German, Pooled or Luxembourgish). Similarities between Luxembourgish and a neighboring language can then be expressed as percentages of aligned segments per language. Even more specifically, we can explore language similarities for individual phonemes or phonemic classes (e.g. unvoiced plosives, diphthongs).

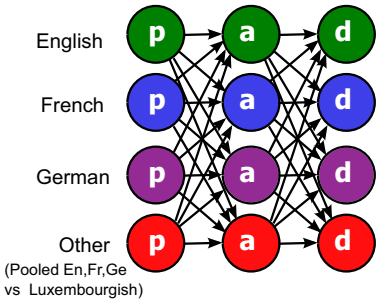


Fig. 1. Multiple choice forced alignment setup: the pronunciation of each word is represented by a fully connected trellis. In the example word *Pad* (Eng.: path), its underlying phonemes (here /pad/) can be realised via either English, French, German or other (supervised pooled vs unsupervised Luxembourgish) models.

5. SIMILARITY RESULTS ON PHONE SEGMENTS

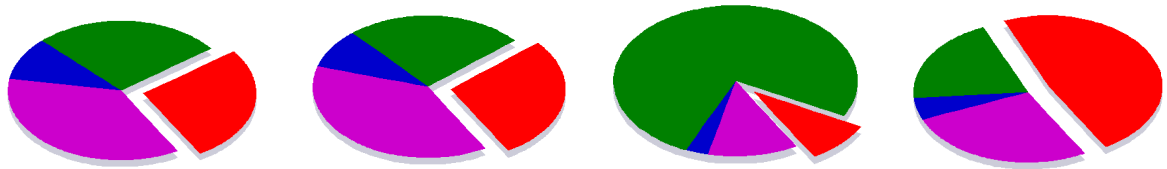
This section reports some of the alignment results.

5.1. Luxembourgish speech data

Forced alignments were carried using 80 minutes of manually transcribed speech from the House of Parliament (*Chamber* debates (70') and from news (10') broadcast by RTL, the Luxembourgish radio and TV broadcast company [2]. The detailed manual transcripts include all audible speech events, including disfluencies and speech errors. These *verbatim* transcripts were checked against the resulting word lists for errors and orthographic inconsistencies. The corpus includes a total of 56,000 phone segments.

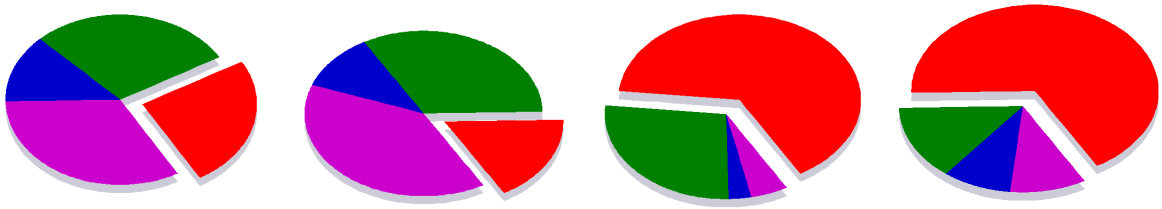
5.2. Results

In earlier investigations [7], a language change in acoustic models was permitted only on word boundaries. In that case, a clear preference for the" German" acoustic models could be noticed with more than 50% of phone segments aligned with the acoustic models arising from the German audio data. As described in section 4, the language identity of the acoustic models may change at any phone segment boundary here. As a result, a large number of language switches between model sets are observed. The German models are less used. However, in all explored alignment conditions (three sets of models trained with monolingual English, French and German data, or adding also a fourth acoustic model set trained with the pooled audio data or the native Luxembourgish models stemming from unsupervised training), the German models are always globally at rank one. In the alignment condition including the pooled model set (see Figure 2 left), German models are aligned with 36% of segments, followed by the English



overall (27 ; 27 ; 10 ; 36) **UV plosives** (28 ; 25 ; 9 ; 38) **/ʊə/** (9 ; 76 ; 3 ; 12) and **/ɪə/** (48 ; 20 ; 5 ; 27)

Fig. 2. Rates of aligned segments per model type (represented as pie charts). The off-set slice corresponds to the **multilingual Pooled model (in red)** followed by English (green), French (blue) and German (magenta) in counter-clockwise order. The different pie charts give from left to right: overall rates, rates for unvoiced plosives, rates for the two diphthongs /ʊə/ and /ɪə/.



overall (25 ; 29 ; 13 ; 33) **UV plosives** (17 ; 33 ; 11 ; 39) **/ʊə/** (65 ; 27 ; 3 ; 5) and **/ɪə/** (67 ; 14 ; 9 ; 10)

Fig. 3. Rates of aligned segments per model type (represented as pie charts). The off-set slice corresponds to the **unsupervised Luxembourgish model (in red)** followed by English (green), French (blue) and German (magenta) in counter-clockwise order. The pie charts give from left to right: overall rates, rates for unvoiced plosives, rates for the two diphthongs /ʊə/ and /ɪə/.

set (27%) and the pooled model set (27%). The French models are used only for 10% of the segments on average. When replacing the pooled acoustic model set by the native Luxembourgish set (see Figure 3 left), overall rates remain very similar (German: 33%, English: 29%), Luxembourgish: 25%, French: 13%). An overall preference for Germanic languages and in particular for the German language can be observed for Luxembourgish speech. This result is in agreement with the postulated influence of typological distance.

To explain the disappointingly low rate for native Luxembourgish models (25%), further investigations are required. However, we may suggest a lower acoustic-phonetic resolution here due to combined methodological and linguistic factors ranging from unsupervised phone labeling and segmentation to high rates of dialectal and idiosyncratic pronunciation variation in Luxembourgish. By contrast, a potential phonetically interesting implication is then that the (supervised) acoustic models may hold across languages for a large range of IPA phone symbols with a relative high cross-language validity. This hypothesis is further explored on a more detailed phonemic basis using the class of unvoiced plosives present in the four considered languages and two (/ʊə/ and /ɪə/) diphthongs which are considered specific for Luxembourgish (not considered to be part of the other languages' phonemic inventories). The achieved rates of the pooled and the native Luxembourgish models will give an indication of the matching accuracy between Luxembourgish speech and the contact languages' monolingual models.

5.2.1. Symbols shared among languages: plosives

The plosives /p/, /t/ and /k/ exist in the four languages. As Luxembourgisch is considered a Germanic language, we may hypothesize that plosives be realized similar to German and English. We may thus expect a stronger burst than in French plosives and positive VOTs. If major acoustic correlates are shared with German and English rather than with French, then plosive segments should be aligned with German or English models rather than with French ones. Detailed results are shown in the second pie charts (from left) of Figures 2 and 3. Results here are very close to the previous overall configuration. Segments are mainly aligned using German and English models. When using the Luxembourgish models instead of the pooled models, the share of the pie charts' off-set slice decreases from 28% to 17%, which can be interpreted as a relatively weak match of the unsupervised native models as compared to the contact languages' supervised models. In the present state, unvoiced plosives are better modelled by the contact languages than by Luxembourgish models.

5.2.2. Approximate symbol mapping: diphthongs

The Luxembourgish phonemic repertoire includes a large number of diphthongs. Here we focus only on the two /ʊə/ and /ɪə/ diphthongs considered as Luxembourgish-specific. Detailed results are shown for /ʊə/ and /ɪə/ in the two pie charts in the right of Figures 2 and 3. In the pooled model condition (Figure 2), either English alone (for /ʊə/) or English and pooled models (for /ɪə/) account for about 75% of the segments. When introducing the Luxembourgish models (Figure 3), the segments corresponding to the two diphthongs are aligned with the Luxembourgish acoustic model for about 65% of the occurrences. This result highlights that the Luxembourgish models do capture acoustic-phonetic specificities quite well and that they are largely chosen, when the corresponding models of the competing languages provide only an approximate match.

Further investigations about the accuracy of the unsupervised acoustic Luxembourgish models are carried out in the next section on automatic speech recognition.

6. AUTOMATIC SPEECH RECOGNITION IN LUXEMBOURGISH

First results of large vocabulary continuous speech recognition (LVCSR) for Luxembourgish were presented in [8] on set of manually transcribed data (70 minutes from CHAMBER and 10 minutes from RTL). The word error rates (WER) were in the range of 55 to 70% on a set of manually transcribed date (70 minutes from CHAMBER and 10 minutes from RTL). These initial recognition experiments supported the observations presented in the previous sections, that the German language provides the closest match to Luxembourgish, followed just behind by the pooled models, then English and finally French. However, in order to obtain recognition word error rates close to those reported for other European languages, it is necessary to estimate the acoustic models on substantially more audio data. Unfortunately, however, no speech corpora with manual transcripts are available for Luxembourgish. Therefore it was decided to apply the semi-supervised acoustic model training developed in [4]. The acoustic models used in the above alignment experiments can serve as initial models for the process, having recognition error rates of the same order of magnitude as the models used to initialize the semi-supervised acoustic model training developed in [4]. The basic idea is to iteratively automatically transcribe a large volume of Luxembourgish speech data, providing indirect supervision via the language model. The transcripts generated in one iteration serve as references for the next, and used for 'pseudo-supervised' acoustic model training. The new models are then used to decode a larger set of audio. At each iteration more accurate acoustic models can be trained (more contexts, more Gaussians). There have been a number of proposed variants for unsupervised training – the number of iterations, the use of filtering or confidence scores, the amount of data used in each iteration is largely empirical [4, 9, 10, 11]. Here we adopt the strategy recently applied in the Quaero project to develop models for the Latvian language [12].

6.1. Speech corpus

As part of the Quaero project we collected a large quantity of speech data in Luxembourgish. It consists of audio data downloaded from the Web in 2012 and 2013, mainly from the RTL channel (flash news, 1000 hours), but also from other sources (Radio100.7, 15 hours; talk shows, 5 hours, ...). In total about 1600 hours of audio were collected and were partly used for unsupervised training of acoustic models in Luxembourgish. The development data (196 min) used to evaluate the ASR is the official Quaero data sets for which manual reference transcriptions were created by the data annotation team. The development data ses comes from the same sources as the those in the training set.

6.2. Language model development

We collected different Luxembourgish texts, some described in [2] and others newly collected from the web. The texts belong to 3 domains:

1. **'New/information' related written sources**:
   - RTL2008: old RTL data (2008 and earlier) manually filtered.
   - RTL2012: Web sites affiliated to RTL (collected in 2012).
   - WIKIPEDIA: Luxembourgish Wikipedia.
   - MISC: miscellaneous reports, books, reviews ... collected on the web.
2. **Oral transcriptions**:
   - CHAMBER: *bona fide* transcriptions [13] of the Luxembourgish Parliament debates.
3. **Social media**:
   - BLOGS: 90 blogs (out of 400 preselected Luxembourgish blogs).
   - BLOGS_COMMENT: user comments from the selected blogs.

6.2.1. Filtering heterogeneous multilingual data

A stochastic language identification system [14] was used to efficiently filter Luxembourgish texts from those in the German, French and English languages in order to process heterogeneous multilingual texts such as are typically harvested from the Web.

The volume of raw texts and of filtered texts are summarized in Table 3. The amount of rejected data (average 33%) strongly depends on the source as expected: for WIKIPEDIA only 3% of the data were rejected[2], while 68% of the Luxembourgish BLOGS were not written in Luxembourgish, according to the automatic identification system, even though only the blogs (90 out of 400) with a significant part of written Luxembourgish were kept. 27% of the CHAMBER texts were also rejected: beyond transcripts in French language due to occasional switches to French language in oral debates, this rather high rejection rate is due to the presence of reports written in French. After filtering, the amount of Luxembourgish-labeled data sums to over 34 Mwords, with an average rejection rate of 33% of the raw texts.

source	size	size	%rejected
RTL2008	611	607	<1
WIKIPEDIA	3603	3483	3
RTL2012	10,307	7948	23
BLOGS_COMMENTS	3106	2386	23
CHAMBER	22,110	16,108	27
MISC	1677	855	49
BLOGS	10,243	3265	68
total	51,657	34,653	33

Table 3. Text size (in thousands of words). (left) Raw texts per data source (7 sources, totaling 51Mwords) (right) Luxembourgish text sizes and percentages of rejected texts.

6.2.2. Effect of filtering on word list and language model

Both the raw and filtered texts were used to build and compare word lists and language models, using the methods described in [2]. The 200k most probable words were selected from the 7 Web data sources, so as to minimize the unigram perplexity. An OOV (Out of Vocabulary) rate of 2.35% was achieved with the filtered sources, to be compared to an OOV rate of 3.23% with the raw texts (28% relative improvement). With respect to the language model, the best interpolated 3-gram model gives a dev set perplexity of 369.35 with the filtered sources (387.20 without filtering, $+5\%$). Due to filtering, the OOV rate exhibits a large improvement, with a more limited gain for the language models. This is generally observed when the amount of texts is insufficient: filtering improves the precision of the word list, however the negative impact on perplexity of filtering out few correct Luxembourgish n-grams counterbalances the positive impact of improving the precision.

[2]some residual non-Luxembourgish languages such as ancient Greek was rejected because of its special coding alphabet

6.3. Acoustic model development for Luxembourgish

The unsupervised acoustic models training process can be summarized as follows:

- Context-independent (CI) acoustic seed models were taken from available sources as seed models. Based on the alignment experiments it was decided to use the pooled acoustic models for initialization.
- The CI acoustic models are used, together with the pronunciation lexicon and language models, to decode a portion of training acoustic data.
- The resulting transcriptions are then used to segment and train a set of context-dependent (CD) acoustic models on the transcribed data.
- These steps are repeated increasing the amount of audio and size of the acoustic models.

The first decoding was carried out using acoustic models trained on pooled data from 3 languages (French, German and English) [8]. In the first 3 iterations, decoding was done with models using plpf0 features, whereas the last decoding was done using models with mlpplpf0 features (the MLP features from the German STT system were used for Luxembourgish). In each iteration the amount of untranscribed data was roughly doubled.

The acoustic models are tied-state, left-to-right 3-state HMMs with Gaussian mixture observation densities (typically 32 components). The triphone-based phone models are word-independent, but position-dependent. The states are tied by means of a decision tree to reduce model size and increase triphone coverage. Since priori experience with unsuperviced acoustic model training for other languages typically had similar results with gender-independent and gender-dependent models, in these experiments only gender-independent models were estimated and tested.

Table 4 summarizes the audio data used in successive acoustic models sets along with the model sizes. The last column reports the word error rate on the Quaero development data set. The upper part of the table gives the initial word error rates for four different sets of context-independent seed models. The WER is about or above 70% for all four sets. Although the German seed models had a slightly lower WER, the multilingual pooled models were used for initialization.

The lower part of the table reports WER with increasingly accurate Luxembourigh acoustic models. The entries two Unsup1 compare the error rates obtained by simply increasing the number of modeled contexts from 7k to 22k, with 80 hours of audio data. The audio data is roughly doubled for Unsup2, with a relative error rate reduction of 11%. A smaller improvement (3%) is observed in the next iteration (Unsup3), and 6% with the MLP parameters for Unsup3. The Unsup3-mlp system was used for our submission in the Quaero 2013

Models	# contexts	#hours	WER(%)
Seed models (Eng)	63	200h	75.6
Seed models (Fre)	63	300h	70.8
Seed models (Ger)	63	52h	61.8
Seed models) (pooled EFG)	63	552h	64.4
Unsup1	7k	80h	33.6
Unsup1	22k	80h	33.2
Unsup2	31k	193h	29.3
Unsup3	39k	500h	28.4
Unsup3, mlp	39k	500h	27.4
Unsup4, mlp	51k	1200h	25.6

Table 4. WER for different acoustic models for Luxembourgish on the Quaero development data set.

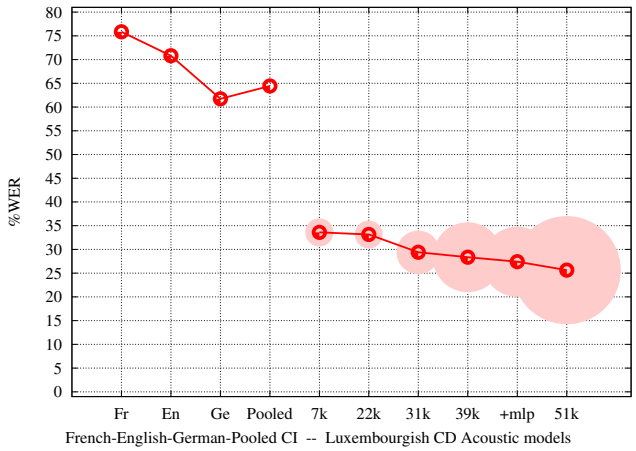


Fig. 4. WER with multilingual exogeneous acoustic models (left) and endogeneous (Luxembourgish) acoustic models (right). The disc around each result represents the speech data training size.

evaluation. After the evaluation, a larger set of 1200 hours of data were transcribed using the Unsup3-mlp models, resulting in a 25.6% WER on the development data.

Figure 4 summarizes the recognition results. The left part of the figure shows the results with the non-native acoustic models (French, English, German and Pooled). The error rates range from 62 to 75%. The right part of the figure shows the WER as a function of the acoustic model size in terms of the number of modeled phone contexts (ranging from 7k to 51k). The light pink circles indicate the size of the audio training corpus, ranging from 80 hours to 1200 hours.

7. SUMMARY AND PROSPECTS

We proposed to study acoustic similarities between Luxembourgish and major contact languages (German, French, English) with the help of automatic speech alignment and recognition systems. The multilingual context in Luxembourg with three official languages (Luxembourgish, German, French) and an increasing influence of English in professional environments offers an interesting testbed to crosslingual and multilingual explorations. The present work focused on the issue of producing acoustic models for automatic speech alignment and recognition in Luxembourgish, a language with strong Germanic and Romance influences. A phonemic inventory was defined and linked to inventories from major neighboring contact languages (German, French and English), using the IPA symbol set. For each of these languages, acoustic seed models were composed using monolingual German, French or English acoustic model sets. Multilingual models were trained using the corresponding pooled audio data and their use during alignments was contrasted with unsupervised native Luxembourgish models. To this end, a multilingual forced alignment process was setup enabling an acoustic model type (of language identity) switch at any phonemic position in pronunciations during the alignment process. The language identity of the aligned acoustic models provides information about the overall acoustic adequacy of both the cross-language phonemic correspondances and the acoustic models. Furthermore, some information can be gleaned on inter-language distances.

We may now come back to our initial research question of how the unsupervised native Luxembourgish models compare to the supervised monolingual seed (German, French, English) models and the pooled models in a multiple choice forced alignment setup. First, in the pooled model condition, it was shown that the Germanic acoustic seed models provided the best match with 36% of the aligned segments using German seeds, 27% using the English ones and another 27% for the pooled models. Only 10% used the French acoustic models. Since Luxembourgish is considered a Western Germanic language close to German, this result is in line with its linguistic typology. When replacing the pooled models by the Luxembourgish ones, the overall aligned segment rate remains almost fixed (dropping slightly from 27% to 25%). This result is surprising and somewhat disappointing: our unsupervised acoustic models do not outperform the models from the other languages on a global basis. However, large differences may be observed depending on phonemic identity. The good news is that the more Luxembourgish-specific the phonemic label, the higher the Luxembourgish alignment rates are. For instance, the Luxembourgish unsupervised models cover only 17% of the unvoiced plosives (vs. 28% for the pooled models), while a strong preference (over 60%) may be noted for the Luxembourgish models of the most specific /ɪə/ and /ʊə/ diphthongs. The fact that models built upon exogeneous multiple languages (some with a high linguistic proximity with Luxembourgish) obtain about

the same coverage as endogeneous models, justify the use of pooled models as seed models for unsupervised training. Futhermore, the pooled model rates give an indicative measure of match/mismatch between the acoustic realisations of a given pair of phonemes between source and target languages. As a perspective, we propose to enhance this measure to provide help for foreign language learning, elaborating lists of potentially difficult phonemes given L1/L2 pairs.

Finally, we addressed the issue of unsupervised native Luxembourgish models performance in automatic speech transcription. ASR results are provided on the official Quaero development set using a language model obtained by filtering the web data with a stochastic language identification system and unsupervised acoustic model training. Our approach accounts for the lack of reliable written resources to develop word lists and language models and takes benefit of the acoustic proximity with other languages to build acoustic seed models for unsupervised training. We have shown that this approach is efficient to build competitive systems for under-described languages. To the best of our knowledge, the final WER of 25.6% which already compares well to other European languages, is the best ever achieved result in Luxembourgish.

8. ACKNOWLEDGEMENTS

This research was funded by the French Oseo Quaero. It was further supported by the French ANR LabEx EFL (ANR/CGI) program.

9. REFERENCES

[1] F. Schanen, *Parlons Luxembourgeois*, L'Harmattan, 2004.

[2] M. Adda-Decker, T. Pellegrini, E. Bilinski, and G. Adda, "Developments of letzebuergesch resources for automatic speech processing and linguistic studies.," in *LREC*, 2008.

[3] C. Krummes, "Sinn si or si si? mobile-n deletion in luxembourgish," in *Papers in Linguistics from the University of Manchester: Proceedings of the 15th Postgraduate Conference in Linguistics*, Manchester, 2006.

[4] L. Lamel, J.L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, no. 1, pp. 115–229, 2002.

[5] P. Gilles and J. Trouvain, "Luxembourgish," *Journal of the International Phonetic Association*, vol. 43, pp. 67–74, 2013.

[6] T. Schultz and A. Waibel, "Experiments on cross-language acoustic modeling," in *Proceedings of Eurospeech*, Aalborg, 2001.

[7] M. Adda-Decker, L. Lamel, and N.D. Snoeren, "Comparing mono- & multilingual acoustic seed models for a low e-resourced language: a case-study of Luxembourgish," in *InterSpeech'10, 11th Annual Conference of the International Speech Communication Association*, Makuhari, Japan, 2010.

[8] Martine Adda-Decker, Lori Lamel, and Gilles Adda, "A first lvcsr system for luxembourgish, an under-resourced european language," in *Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics (L&TC 2011)*, Poznan, Poland, 25/11 au 27/11 2011, pp. 47–50.

[9] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," in *IEEE Transactions on Speech and Audio Processing*, 2005, pp. 23–31.

[10] J. Ma and R. Schwartz, "Unsupervised versus supervised training of acoustic models," in *Interspeech 2008*, 2008, pp. 2374–2377.

[11] Scott Novotney, Richard Schwartz, and Jeff Ma, "Unsupervised acoustic and language model training with small amounts of labelled data," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4297–4300.

[12] Ilya Oparin, Lori Lamel, and Jean-Luc Gauvain, "Rapid development of a latvian speech-to-text system.," in *ICASSP*. 2013, pp. 7309–7313, IEEE.

[13] Martine Adda-Decker, Claude Barras, Gilles Adda, Patrick Paroubek, Philippe Boula de Mareïl, and Benoit Habert, "Annotation and analysis of overlapping speech in political interviews.," in *LREC*. 2008, European Language Resources Association.

[14] Thomas Lavergne, Olivier Cappé, and François Yvon, "Practical very large scale CRFs," in *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*. July 2010, pp. 504–513, Association for Computational Linguistics.