Comparing mono- & multilingual acoustic seed models for a low e-resourced language: a case-study of Luxembourgish

Martine Adda-Decker, Lori Lamel, Natalie D. Snoeren

Spoken Language Processing Group, LIMSI-CNRS, 91403 Orsay, France

madda@limsi.fr, lamel@limsi.fr, nsnoeren@limsi.fr

10.21437/Interspeech.2010-108

Abstract

Luxembourgish is embedded in a multilingual context on the divide between Romance and Germanic cultures and has often been viewed as one of Europe's under-resourced languages. We focus on the acoustic modeling of Luxembourgish. By taking advantage of monolingual acoustic seeds selected from German, French or English model sets via IPA symbol correspondances, we investigated whether Luxembourgish spoken words were globally better represented by one of these languages. Although speech in Luxembourgish is frequently interspersed with French words, forced alignments on these data showed a clear preference for Germanic acoustic models with only a limited usage of French. German models provided the best match with 54% of the data, 35% for English and only 11% for French models. A set of multilingual acoustic models, estimated the pooled German, French, and English audio data, captured 27% to 48% of the data depending on conditions.

Index Terms: multilingual alignment, acoustic seed models, under-resourced languages, Luxembourgish, English, French, German.

1. Introduction

Luxembourg, a small country of less than 500,000 inhabitants in the center of Western Europe, is composed of about 65% of native inhabitants and 35% of immigrants. The national language, Luxembourgish ("Lëtzebuergesch"), is considered as an official language of Luxembourg only since 1984 and spoken by natives, with French and German being easily used for communication among residents [1]. The immigrant population generally speaks or learns one of Luxembourg's other official languages: French or German. Recently, English has joined the set of prestigious languages of communication, and tends to become a major communication tool in professional environments.

As pointed out by [2] and [3], Luxembourgish should be considered as a partially under-resourced language, mainly because of the fact that the written production remains relatively low and that linguistic knowledge and resources, such as lexica and pronunciation dictionaries, are sparse. Rather surprisingly, written Luxembourgish is not systematically taught to children in primary school: German is usually the first written language learned, followed by French. Even today, German and French are the most practiced languages for written communication and administrative purposes in Luxembourg, guaranteeing a larger dissemination, whereas Luxembourgish is mainly being used for oral communication. The question then arises, whether the acoustic realization of Luxembourgish phonemes are mainly influenced by German or French, or even by the less practiced English language. It is difficult to estimate the numbers of Romance/Germanic influenced words in Luxembourgish, as proportions greatly depend on communicative settings. Vernacular Luxembourgish is highly influenced by its Germanic filiation, whereas more technical and administrative jargons include a higher proportion of Romance words. A further question then concerns whether French related words are better represented by French acoustic models.

The goal of this paper is to gain more insight into the acoustic properties that define the Luxembourgish language in the light of its Germanic and Romance influences. We focus on the acoustic modelling and on the elaboration of phonemically aligned audio data. The following questions were addressed. First, when aligning Luxembourgish audio data using monolingual acoustic seeds in parallel for several languages, are the language-specific seeds randomly used or is there a clear preference for one language? Second, is there a language-specific preference to be observed for specific phonemes or for phoneme sets? If so, do they correspond to IPA symbol matches between the preferred language and Luxembourgish phonemes? Third, how do language-specific preferences, if any, fare with added pooled multilingual acoustic phone models? The raised issues have important implications for ASR studies of acoustic modelling and the processing of pronunciation variants. The next section introduces the phonemic inventory of Luxembourgish and its correspondance with the three source languages (German, French, and English). Next, the alignment results are presented with both monolingual and multilingual seed models. Finally, section 3 provides a summary of the results and discusses some major future challenges for speech technology and linguistic studies of Luxembourgish.

2. Phonemic inventory and alignment

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 [4] 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 /a/ and /j/ for diphthong a͜ɪ). We prefered, 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.

Alignment experiments were carried out using different initializations for the Luxembourgish acoustic models and different pronunciation dictionaries. To this end, we manually transcribed 80 minutes of speech from the House of Parliament (*Chamber* debates and to some extent from news channels, delivered by the Luxembourgish radio and televion broadcast company RTL (see [2] for more information on the Luxembourgish corpora that are currently available).

2.1. Acoustic seed models

The need for the development of acoustic seed models for underresourced languages has already been addressed in previous research [5]. In the current study, three sets of context-

Table 1: Excerpts from the cross-lingual phone association table, the Luxembourgish pronunciation dictionary and the multilingual dictionary used for alignments with the multilingual acoustic super-set. Luxembourgish target phonemes are associated to identical or similar (in grey) phonemes of the different source languages (French, German, English).

Carrier word (Eng)	Lux	Fre	Ger	Eng
ORAL VOWELS				
l**ii**cht (light)	i	i	i	i
sch**ü**tzen (shelter)	ʏ	y	ʏ	ɪ
f**ä**eg (able)	ɛ:	ɛ	ɛ:	ɛ
DIPHTHONGS				
l**éi**en (to tell lies)	e͜ɪ	e	e	e
l**ou**nen (to hire)	ɔ͜ʊ	o	o	o
FRICATIVES & AFFRICATES				
Ee**ch**en (oak tree)	ç	ʃ	ç	ʃ
Li**g**en (lie)	ʝ	ʒ	ç	ʒ
NASALS & GLIDES				
Hä**r** (mister)	ɐ̯	ə	ɐ̯	ə

Table 2: Phoneme and training information for the native and the pseudo-Luxembourgish acoustic models (using either English, French or German acoustic model sets) and the super-set of multilingual acoustic seeds.

	#native		*#Luxemb.*
Language	*phonemes*	*#training (h)*	*phonemes*
English	48	150	60
French	37	150	60
German	49	40	60
Superset(E,F,G)	-	340	3x60

independent and gender-independent acoustic models were built, one for each seed language (i.e., English, French, German). The models were trained on manually transcribed audio data (between 40 and 150 hours) from a variety of 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 is a tied-state left-to-right, 3-state CDHMM with Gaussian mixture observation densities (typically containing 64 components). The acoustic features are derived from a PLP-like [6] acoustic parametrization, which has been used in the LIMSI systems since 1996 [7]. The coefficients are normalized on a segment cluster basis using cepstral mean removal and variance normalization. The resulting cepstral coefficient for each cluster has a zero mean and unity variance. The 39-component acoustic feature vector consists of 12 cepstrum coefficients and the log energy, along with the first and second order derivatives.

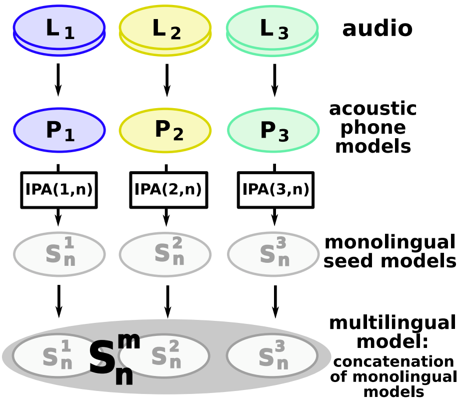

Figure 1: Monolingual S_n^i and multilingual S_n^m acoustic seed models for a new language n (Luxembourgish) given phone models P_i of languages L$_i$ ($i = 1, 2, 3$ English, French, German; $m = 1 + 2 + 3$) and IPA symbol correspondances between language i and n IPA(i,n).

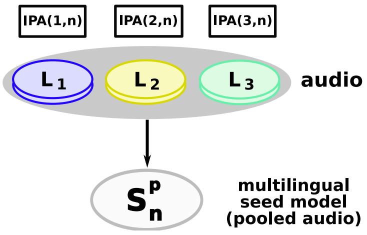

Figure 2: Pooled multilingual S_n^p acoustic seed model for a new language n (Luxembourgish) given transcribed audio data of languages L$_i$, ($i = 1, 2, 3$, English, French, German) and IPA symbol correspondances between languages i and n IPA(i,n).

Figure 1 illustrates the development of four sets of pseudo-Luxembourgish acoustic models, each including 60 phones, starting from the English, French and German seed models and mapping the Luxembourgish phonemes to a close equivalent in each of the source model sets (IPA(i,n) in Fig. 1). Table 1 shows a sample of the adopted cross-lingual associations, to initialize seed models for Luxembourgish. Some symbols are used several times for different Luxembourgish phonemes. In particular, for the diphthongs that are missing in French, we chose to select the phonemes corresponding to the nucleus vowel. A fourth model set was formed by concatentating the first three model sets, allowing the decoder to choose among the three models (see Table 2). Finally a set of multiligual acoustic models were trained using the pooled E,F,G audio data that were labeled using their respective IPA(i,n) correspondances (see Fig. 2).

Table 3: Excerpt of the Luxembourgish pronunciation dictionary used for alignment. Upper part: examples of standard pronunciation. Lower part: excerpts of the pronunciation dictionary for alignments with the multilingual acoustic super-set.

PRONUNCIATION DICTIONARY	
lexical entry (English)	*citation form*
déi (those) Kapp (head)	de͜ɪ kʌp
MULTILINGUAL DICTIONARY	
déi (those)	d$_g$e͜ɪ$_g$ d$_f$e͜ɪ$_f$ d$_e$e͜ɪ$_e$ d$_g$e͜ɪ$_e$ d$_f$e͜ɪ$_g$ d$_f$e͜ɪ$_e$ ...

2.2. Luxembourgish audio alignment

For the alignment experiments, we employed the Luxembourgish audio corpus with corresponding detailed acoustic transcripts, comprising a total of 80 minutes of manually transcribed audio data (Chamber (70') and News (10')). The detailed transcripts were generated from scratch for the news data. For the *Chamber* data, the audio stream was manually segmented into speaker turns, according to the existing *bona fide* report. For each speaker, the *bona fide* transcriptions were changed if necessary to faithfully reflect the speech flow. All uttered audible speech events, including disfluencies and speech errors, were manually transcribed. The quality of the manual *verbatim* transcripts were checked against the resulting word lists for errors and orthographic inconsistencies.

Table 4: Average proportions of aligned German, English, French seeds in the multilingual super-set configuration. Results are given for a subset of selected phonemes.

Phone type	*German*	*English*	*French*	# occ.
overall	54.3	35.3	10.4	55873
p	67.05	21.85	11.10	865
t	55.91	35.23	8.86	3588
k	55.15	36.64	8.21	1048
ç	56.80	34.52	8.67	588
χ	80.87	14.29	4.84	413
h	36.05	59.36	4.59	785
ʒ	41.96	25.00	33.04	112
y	25.00	15.62	59.38	32
ʏ	41.03	25.64	33.33	39

Speech parts which poorly match their transcripts (i.e. the corresponding acoustic models) as evidenced by a temporal mismatch (abnormally long segments) are rejected. This is controlled by a duration threshold which is the same for all languages. The English language exhibited the highest rejection rate (i.e. 516s, corresponding to 10% of the data), whereas rejection rates were much lower for the other settings, the lowest rates being with the German language (131 s, < 3% (see [4] for more information with respect to rejections).

The average phone segment duration remains relatively stable with respect to the different monolingual seed alignments. Variations here stem from variable proportions of the acoustic signal assigned to the extra-phonemic models. The German alignment yields the smallest phone duration of 0.07s on average (silence, breath and hesitation segments not being considered). For English and French, the average segment duration amounts to 0.08s. It can be seen that, independent of acoustic-phonetic considerations, the German silence (including background noise) model was used more frequently during the German monolingual alignment, than was the case for the French or English silence models. This explains the smaller average phone duration, and could be related to the relatively small volume of training data (40h) for the German originated seeds (as opposed to French and English), with a lower capacity to cover various acoustic conditions.

2.3. Multilingual alignments

The alignment produced by the acoustic super-set model, together with the multilingual pronunciation dictionary achieves the highest proportion of aligned acoustic phone segments. In this configuration, it is interesting to assess the models at two levels. First, on the phone segment level, one can measure the proportions of segments aligned using the seeds of a given language. Are there differences in proportions as a function of specific phonemes? Second, on the word level, one may want to check whether the proportion of aligned French seeds is higher for French loan words than for native Luxembourgish words. For instance, one might expect that for Luxembourgish diphthongs, the proportion of aligned English seeds may increase, especially for diphthongs not covered by the German language. Conversely, the proportion of French and English seeds used for Luxembourgish and German specific sounds (e.g. χ) should remain very low. Table 4 displays aligned monolingual seed proportions as produced by the multilingual super-set. More than half of the 55873 segments were aligned using the German seeds. About one third corresponds to English seed models and only 10% of the segments were aligned using the French models. Results for some phonemes are shown to illustrate that proportions can notably vary across phoneme identity.

We began the present study by asking ourselves whether a preference for a particular language can be observed when using monolingual acoustic seed models. Our results suggest that the answer to this question depends on the alignment level (phoneme or word). When there is a possible language switch at word boundaries only, German achieves the highest acoustic model selection rates for almost all consonants (except for /h/, /ŋ/ and post-vocalic “r” /ɐ̯/ that have a majority of English model assignments) as well as vowels. Concerning the latter, the English models are slightly preferred for /ʌ/, /ɛː/, /ʊə̯/ and /ɛɪ̯/. French is always rarely selected when possible language switches take place at the level of word boundaries. These results are highlighted in Figure 3 (left) and may indicate some acoustic channel mismatch, beyond acoustic-phonetic differences between French and Luxembourgish phonemes. The alignments that allow language switches at the phone boundary level enabled us to investigate the acoustic channel mismatch (see Figure 3, right-hand panel). From these graphs, it can be seen that the German dominant profile is somewhat attenuated. That is, more English and even some French phonemes appear to become dominant without the monolingual acoustic word model constraint.

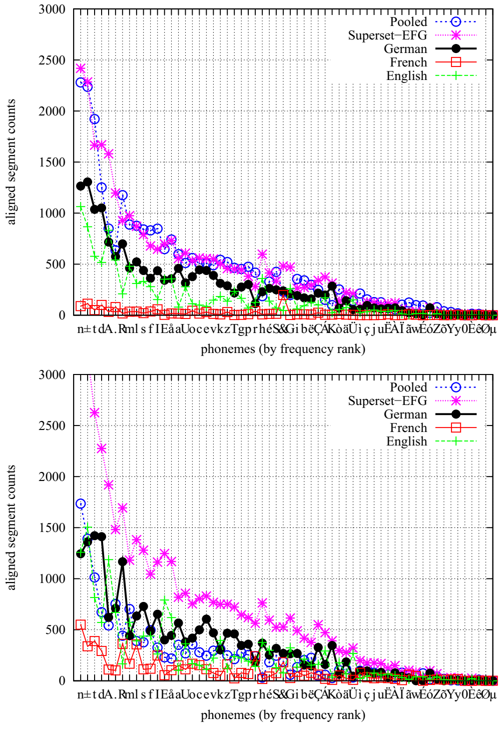

Figure 4: Model (superset of monolingual sets vs pooled) selection rates with align. protocol 1 (model switch on word boundaries): **top**; align. protocol 2 (model switch on phone boundaries): **bottom**.

With regard to the second question as to whether there are language preferences for specific phonemes and/or phoneme sets, our initial results suggest that the English acoustic seeds provide the best match with diphthongs, whereas the French acoustic seeds provide the best match with nasal vowels (although there are relatively few occurrences). These observations obviously call for a more in-depth analysis using a larger corpus. Finally, we addressed the question of having pooled multilingual seed models. Our results indicate that using a super-set model, the German acoustic seeds played a dominant role. About a half of all segments were aligned properly using German models, followed by English and French.

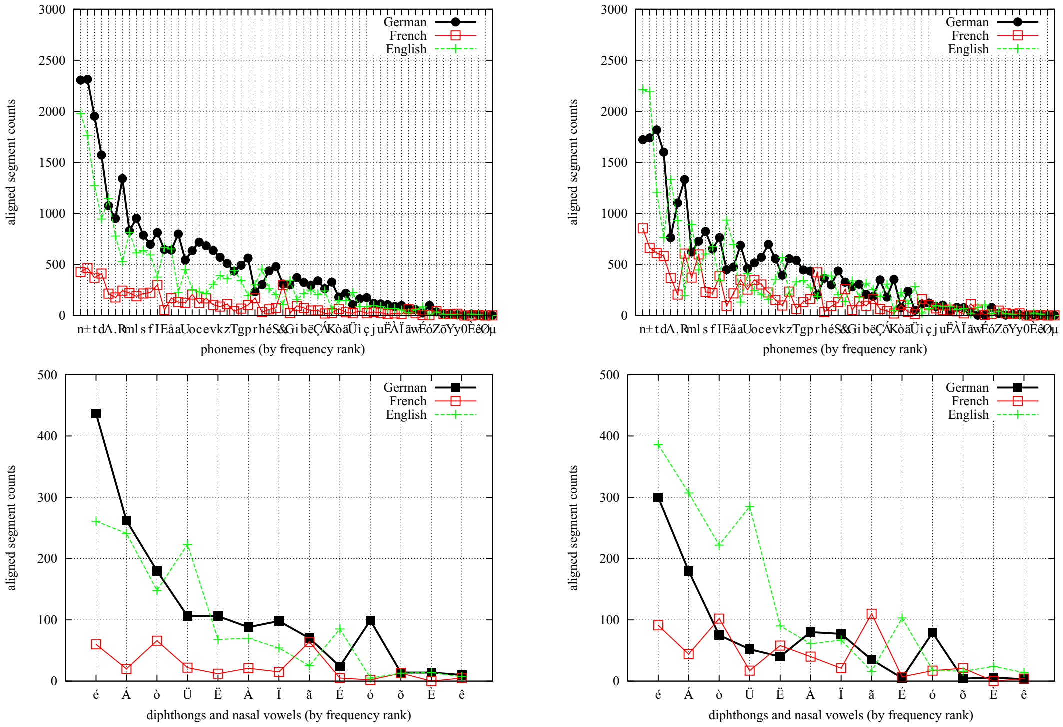

Figure 3: Alignment protocol 1: language switch licensed on word boundaries (left) vs protocol 2: language switch licensed on phone boundaries (right). Model selection rates are shown for all phonemes (top); for diphthong and nasal vowels (bottom).

3. Summary and prospects

The main goal of the present contribution was to draw attention to the complex linguistic situation of Luxembourgish, a partially under-resourced and under-described language. In the present work, we focused on the issue of producing acoustic seed models for Luxembourgish, a language with strong Germanic and Romance influences. A phonemic inventory was defined and linked to inventories from major neighboring languages (German, French and English), using the IPA symbol set. For each of these languages, acoustic seed models were composed using either monolingual German, French or English acoustic model sets. In Luxembourgish speech alignments, a super-set of multilingual acoustic seeds was used putting together the three language-dependent sets. 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. It was shown that the German acoustic seed models provided the best match with 54.3% of the segments aligned using German seeds, 35.3% using the English ones and only 10.4% using the French acoustic models. Since Luxembourgish is considered a Western Germanic language close to German, this result is in line with its linguistic typology.

Computational ASR investigations and corpus-based analyses will not only enhance the development of a more full-fledged ASR system for Luxembourgish, but can also be used to generate more specific predictions about lexical processing in human listeners. For instance, an interesting implication of the present study pertains to the role of previous language knowlegde in foreign language learning and the perception of novel phonemic contrasts. The alignment results suggest that the influence of language-specific phonemic inventories in alignment varies as a function of the locus of language-switching (i.e. at word boundaries or phone boundaries). This prediction could be empirically tested with second-language learners of Luxembourgish. Given the implications of large corpus-based analyses, it is hoped that this line of research on Luxembourgish will sparkle more interest for this language in researchers working in the domains of ASR and linguistics.

4. Acknowledgements

This work has been partially financed by OSEO under the QUAERO program. Part of the research has been supported by a Luxembourgish FNR grant awarded to N.D.S.

5. References

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

[2] M. Adda-Decker, T. Pellegrini, E. Bilinski, and G. Adda, "Developments of lëtzebuergesch 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] M. Adda-Decker, L. Lamel, and N. Snoeren, "Initializing acoustic phone models of under-resourced languages: a case-study of luxembourgish," in *SLTU*, 2010.

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

[6] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech.," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[7] J.L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, vol. 37, pp. 89–108, 2002.