Modélisation acoustico-phonétique de langues peu dotées : Études phonétiques et travaux de reconnaissance automatique en luxembourgois

Martine Adda-Decker[1, 2] Lori Lamel[2] Gilles Adda[2]
(1) LPP, CNRS-Paris 3/Sorbonne Nouvelle
(2) Groupe TLP, LIMSI-CNRS
martine.adda-decker@univ-paris3.fr, lori.lamel@limsi.fr ,gilles.adda@limsi.fr

RÉSUMÉ

Le luxembourgeois est une langue germano-franconique et l'une des langues européennes sous-décrites. Cet article étudie la similitude entre les segments phonétiques en luxembourgeois avec leurs équivalents en allemand, français et anglais *via* des techniques d'alignement forcés. En utilisant les modèles acoustiques monolingues d'amorçage de ces trois langues, ainsi que des modèles "multilingues" entraînés sur un corpus de parole obtenu par concaténation, nous avons examiné si le luxembourgeois était mieux représenté par l'une des langues prises individuellement ou par le modèle multilingue. Au niveau global, les modèles allemands fournissent la meilleure correspondance, mais une analyse par segments montre des préférences spécifiques. Les premiers résultats en transcriptions illustrent les performances des différents jeux de modèles acoustiques monolingues et multilingues, ainsi que les modèles luxembourgeois construits à partir de 1200 heures de parole non transcrites en luxembourgeois, et des méthodes non supervisées.

ABSTRACT

Acoustic-phonetic modeling for under-resourced languages : an overview of recent phonetic studies and automatic speech recognition experiments in Luxembourgish

Luxembourgish, a Germanic-Franconian language, is embedded in a multilingual context on the divide between Romance and Germanic cultures and remains one of Europe's under-described languages. This paper investigates the similarity between Luxembourgish phone segments with German, French and English via forced speech alignment techniques. Making use of monolingual acoustic seed models from these three languages, as well as "multilingual" models trained on pooled speech data we investigated whether Luxembourgish was globally better represented by one of the individual languages or by the multilingual model. While globally, the German models provide the best match, a phone-based analysis, shows language-specific preferences. First ASR results illustrate the accuracy of the various sets of monolingual and multilingual acoustic models and Luxemburgish acoustic models built from 1200 hours of untranscribed Luxemburgish audio data using unsupervised methods.

MOTS-CLÉS : Langues peu dotées ; modélisation acoustique ; modèles multilingues ; système de transcription de la parole ; luxembourgeois ; alignements forcés..

KEYWORDS: under-resourced languages ; acoustic modeling ; multilingual models ; large vocabulary speech recognition ; Luxembourgish ; Forced alignment..

1 Introduction

Le Luxembourg est un pays au centre de l'Europe de l'Ouest, composé de 65% d'habitants autochtones et de 35% d'immigrés. La langue nationale, le luxembourgeois ("Lëtzebuergesch"), n'est la langue officielle que depuis 1984 et n'est parlé que par les autochtones ; les immigrés selon leur pays d'origine parlent l'une des deux autres langues officielles, le français et l'allemand. A ces langues officielles s'est ajouté il y a peu l'anglais en tant que langue de communication fréquente, en particulier dans les milieux professionnels (banque, commission européenne). Le luxembourgeois peut être considéré comme une langue partiellement peu dotée (Adda-Decker *et al.*, 2008b). En effet, il y a peu ressources linguistiques telles que des lexiques ou des corpus de langue écrite en luxembourgeois au profit des écrits en français et en allemand.

Dans ce travail, nous allons examiner les propriétés acoustiques qui définissent la langue luxembourgeoise dans sa relation avec ses influences germaniques et romanes. Nous regardons en particulier l'influence de l'allemand, le français et l'anglais sur la réalisation acoustique des phonèmes du luxembourgeois. Pour cela nous utiliserons des données audio alignées au niveau phonémique. Nous examinerons ensuite comment on peut utiliser cette connaissance pour construire des modèles d'amorçage pour produire des modèles du luxembourgeois en utilisant un apprentissage non-supervisé.

La section suivante présente l'inventaire phonémique du luxembourgeois et sa correspondance avec les trois langues exogènes présentes aux Luxembourg. Nous présentons des résultats d'alignement pour des modèles d'amorçage acoustiques mono- et multilingues. Ensuite nous présentons les résultats en reconnaissance utilisant des modèles acoustiques issus de ces modèles d'amorçage, et appris à l'aide de méthodes non-supervisées sur un grande quantité de données en luxembourgeois. Enfin, nous concluons et présentons les perspectives tant en transcription que pour les études linguistiques sur le luxembourgeois.

2 Similarité entre segments phonémiques

2.1 Inventaire phonémique du luxembourgeois

L'inventaire phonémique du luxembourgeois que nous avons adopté (Schanen, 2004) contient 60 symboles dont 3 symboles extra-phonémiques pour le silence, la respiration et l'hésitation. La Table 1 présente un échantillon de l'inventaire phonémique avec des exemples. Le luxembourgeois est caractérisé par un grand nombre de diphtongues. Nous avons choisi de coder les diphtongues et les affriqués à l'aide d'un seul symbole. Étant donnée l'importance de l'apport du français, nous avons inclus les nasales dans l'inventaire, bien qu'elles ne soient théoriquement pas présentes dans les mots luxembourgeois.

2.2 Modèles acoustiques d'amorçage

Nous avons construit 3 jeux de 60 modèles acoustiques suivant le travail initié dans (Adda-Decker *et al.*, 2010, 2011b), en utilisant les langues exogènes présentes au Luxembourg et pour lesquelles il existe une quantité suffisante de données d'apprentissage. Ainsi, nous avons utilisé 150 heures

exemple (Français)	Lux	Fra	All	Ang
VOYELLES ORALES				
li**i**cht (lumière)	i	i	i	i
sch**ü**tzen (bus)	ʏ	y	ʏ	ɪ
f**ä**eg (capable)	ɛ :	ɛ	ɛ :	ɛ
DIPHTONGUES				
l**éi**en (mentir)	e͡ɪ	e	e	e
l**ou**nen (louer)	ɔ͡ʊ	o	o	o

TABLE 1 – Echantillon de correspondances cross-lingues entre phonèmes : Les cibles en luxembourgeois sont mises en correspondance avec un phonème identique ou similaire dans les 3 langues exogènes (Fra, All, Ang).

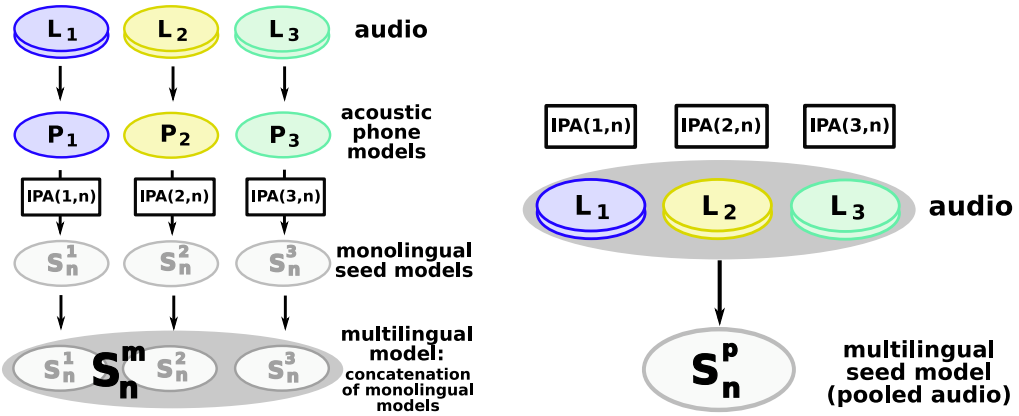

FIGURE 1 – Modèles acoustique d'amorçage pour la langue cible n (luxembourgeois) étant donnés des modèles de phones P_i d'une langue L_i ($i = 1, 2, 3$ anglais, français, allemand), et les mises en correspondance des symboles IPA entre les langues i et n IPA(i,n). gauche : monolingue S_n^i et multilingue par concaténation des modèles S_n^m ; droite : modèles multilingues par concaténation des données S_n^p .

pour l'anglais et le français, et 40 heures pour l'allemand. Les jeux de phonèmes monolingues sont de taille différente : 48 pour l'anglais, 37 pour le français et 49 pour l'allemand. Les modèles de phones sont des modèles de Markov cachés représentant des allophones contextuels avec une structure gauche-droite à états liés utilisant des mélanges de gaussiennes (typiquement 64).

La figure 1 (gauche) illustre le développement de 3 modèles acoustiques de pseudo-luxembourgeois, contenant chacun 60 phones, pour l'anglais, le français et l'allemand, en mettant en correspondance les phonèmes du luxembourgeois avec leur plus proche équivalent dans la langue source (IPA(i, n) dans la Fig. 1). Un quatrième modèle a été construit en concaténant les 3 modèles précédents, en laissant au décodeur le choix parmi ceux-ci. Enfin un jeu de modèles multilingues a été appris en mettant en commun les trois corpus d'apprentissage (figure 1 droite, puis dénommé *pooled* dans le reste de l'article), en utilisant les labels fournis par les correspondances IPA(i, n). Ces jeux multilingues reprennent des concepts introduits par exemple dans (Schultz et Waibel, 2001).

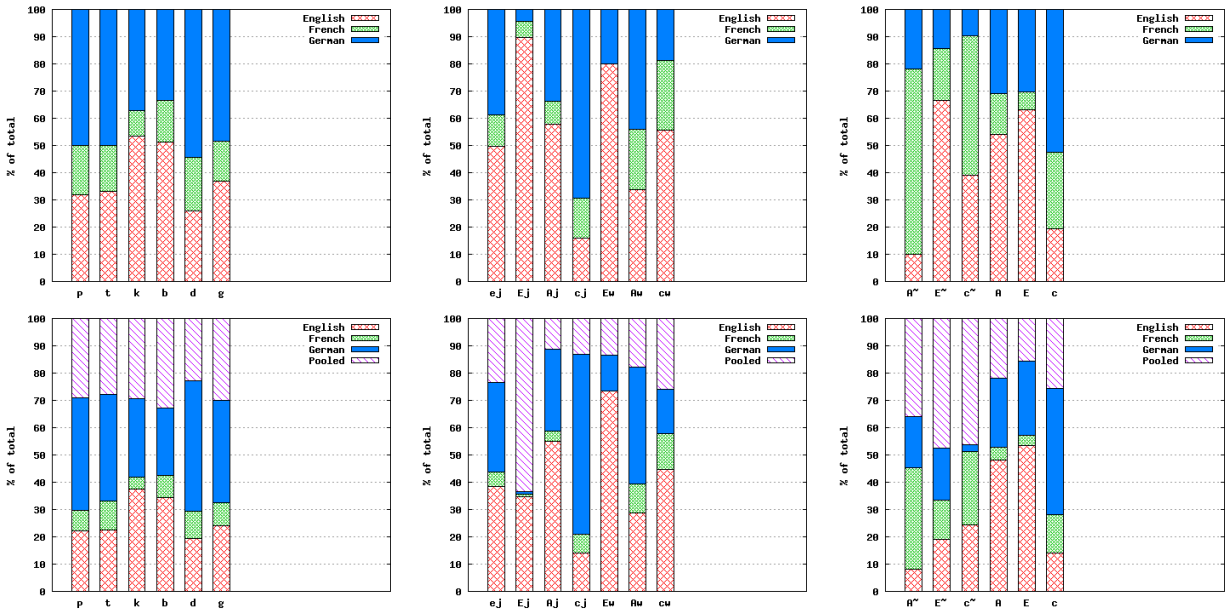

FIGURE 2 – Alignement en utilisant les modèles multilingues par concaténation des modèles monolingues (haut) et en ajoutant le modèle multilingue obtenu par concaténation des données d'apprentissage (pooled, bas). Les taux de sélection sont montrés pour chaque phonème ; gauche : plosives (/p,t,k,b,d,g/) ; milieu : diphtongues /e^j,ɛ^j,a^j,ɔ^j,ɛ^w,a ^{w},ɔ^w/ ; droite : nasales /ɛ̃ ,ã,õ/ et les voyelles orales correspondantes /ɛ,a,ɔ/.

2.3 Données de parole en luxembourgeois pour l'alignement forcé

Les alignements forcés ont été réalisés sur 80 minutes de parole transcrites manuellement, en provenance du parlement luxembourgeois (débats de la Chambre (70')) et des actualités en provenance de la chaîne RTL (10') (Adda-Decker *et al.*, 2008b).

La durée moyenne d'un phone est assez stable (70-80ms) quels que soient les modèles d'amorçage utilisés. Le corpus contient un total de 56 000 segments de phones.

2.4 Résultats

Dans cette étude où nous désirons examiner les propriétés phonémiques des différents modèles, l'identité de chaque langue peut changer à la frontière de chaque phone. De ce fait, on observe de nombreux changements de langue et les modèles allemands sont moins utilisés que dans une étude précédente (Adda-Decker *et al.*, 2010), où le changement avait lieu à chaque frontière de mot. Dans l'alignement avec les modèles obtenus par concaténation des données, les modèles allemands sont alignés avec 36% des segments, les modèles anglais avec 27%, et les modèles français uniquement 10,5%. Dans toutes les conditions, les langues germaniques sont préférées, ce qui est compatible avec un influence corrélée à une distance typologique.

Dans la mise en correspondance entre les symboles IPA du luxembourgeois et des langues exogènes, il y a des symboles qui sont partagés (les plosives) et d'autres pour lesquels la correspondance est approximative (les diphtongues). Nous donnons ici (voir figure 2) quelques détails sur les résultats d'alignement en fonction de la proximité des symboles lors de la mise en

correspondance. Pour chaque symbole, les segments correspondants sont alignés avec l'un des modèles acoustiques (anglais, français, allemand, *pooled*). Les taux sont calculés en fonction de la langue d'origine choisie. Le taux correspondant au modèle *pooled* donne une indication de l'écart entre la langue cible et les modèles monolingues sources.

(i) symboles partagés entre les langues : les plosives Les plosives /p/, /t/ and /k/ et leur correspondant voisé existent dans les 4 langues (sources et cible). Le luxembourgeois étant considéré comme une langue germanique, on peut faire l'hypothèse que les plosives doivent être réalisées d'une façon similaire à celle de l'allemand ou de l'anglais. Les résultats détaillés dans la figure 2 (gauche) montrent que les segments sont partagés entre l'allemand et l'anglais et que seuls 10 à 20% des segments utilisent le modèle français. En ajoutant les modèles *pooled* (bas), environ 30% des données utilisent ce modèle.

(ii) correspondance approximative : les diphtongues Le répertoire phonémique du luxembourgeois contient un grand nombre de diphtongues. La figure 2 (milieu) montre un tableau moins homogène que pour les plosives. Il y a une tendance à utiliser les modèles anglais, les segments du phonème /$\underset{\smile}{\varepsilon\text{ɪ}}$ / étant alignés à 90% avec des modèles anglais, alors que /$\underset{\smile}{\text{ɔ}\text{ɪ}}$ / sont clairement plus allemand. Les modèles *pooled* semble absorber /$\underset{\smile}{\varepsilon\text{ɪ}}$ /, alors que la situation reste inchangée pour /$\underset{\smile}{\text{ɔ}\text{ɪ}}$ /.

(iii) correspondance approximative : les voyelles nasales Les voyelles nasales sont utilisées en luxembourgeois pour certains mots importés du français. La figure 2 (droite) montrent les résultats pour les nasales et les voyelles orales correspondantes. Le taux d'utilisation du modèle français est très élevé pour les segments ã et õ. En introduisant les modèles *pooled* ce taux tend à chuter ce qui montre le faible lien entre les nasales du français et du luxembourgeois.

On peut voir dans la figure 2 (bas) que le modèle *pooled* est utilisé pour aligner un grand nombre de segments (par exemple /$\underset{\smile}{\varepsilon\text{ɪ}}$ /) pour certains phonèmes, alors que pour d'autres le modèle *pooled* n'est utilisé que pour une faible partie des données.

3 Application à la transcription automatique du luxembourgeois

Des expériences préliminaires (Adda-Decker *et al.*, 2011a) nous ont permis de voir les limites de modèles multilingues tels que développés dans la section précédente. Afin d'atteindre des taux d'erreurs acceptables et en l'absence de transcription de corpus audio en luxembourgeois, les modèles développés dans la section précédentes doivent servir de modèles d'amorçage pour la construction de modèles acoustiques en luxembourgeois, en utilisant la méthodologie développée dans (Lamel *et al.*, 2002) : une grande quantité de données audio en luxembourgeois est décodée de manière itérative à l'aide de modèles de langage en luxembourgeois et des modèles acoustiques obtenus lors de la précédente itération. A chaque itération, et au fur et à mesure que les modèles acoustiques sont plus précis et plus efficaces, une plus grande quantité de données audio et un plus grand nombre de contextes phonétiques sont utilisés. Le nombre l'itération nécessaires et la

quantité de données audio incorporée à chaque itération est largement empirique, et la solution retenue ici correspond à la méthodologie utilisée dans (Oparin *et al.*, 2013).

3.1 Donnée audio en luxembourgeois

Nous avons recueilli dans le cadre du projet Quaero une grande quantité de données audio en luxembourgeois. Elles correspondent à des données audio recueillies sur le Web en 2012 et 2013, principalement en provenance de RTL (flash 1000h, journal 880h) mais également d'autres origine (Radio100.7 15h, talk show 5h,...). Le nombre total d'heures récoltées est de 1678. Ces données sont utilisées en partie pour l'apprentissage non-supervisé des modèles acoustiques en luxembourgeois. Les jeux de développement (196mn) et d'évaluation (205mn) sont les jeux officiels fournis dans le cadre du projet Quaero, provenant des mêmes sources que celles présentes dans l'apprentissage.

3.2 Modèle de langage

Nous avons utilisé un système d'identification de la langue (Lavergne *et al.*, 2010) afin de filtrer de manière efficace le luxembourgeois des langues exogènes mentionnées précédemment, afin de pouvoir utiliser les données hétérogènes recueillies sur le Web.

Utilisation de textes multilingues hétérogènes Nous avons utilisé divers textes en luxembourgeois, certains décrits dans (Adda-Decker *et al.*, 2008b) et d'autres recueillis plus récemment sur le Web (Adda-Decker *et al.*, 2011a). Les textes appartiennent à 3 domaines :

1. **Actualités, sources écrites** :
   - RTL2008 : données RTL anciennes (avant 2008) et filtrées manuellement (0,6Mmots)
   - RTL2012 : Sites Web de RTL (aspirés en 2012) (10,3Mmots).
   - WIKIPEDIA : le Wikipedia luxembourgeois (3,6Mmots).
   - MISC : des rapports, livres, revues. . .recueillis sur le Web (1,7Mmots).
2. **Transcription de l'oral** :
   - CHAMBRE : transcriptions *bona fide* (Adda-Decker *et al.*, 2008a) des débats au parlement luxembourgeois (22,1Mmots).
3. **média sociaux** :
   - BLOGS : 90 blogs (provenant d'une liste de 400 blogs en répertorié comme luxembourgeois) (10,2Mmots)
   - BLOGS_COMMENT : les commentaires des internautes sur les blogs préselectionnés (3,1Mmots).

En utilisant le système de détection, on filtre en moyenne 33% des données. La quantité rejetée dépend fortement de la source : pour WIKIPEDIA uniquement 3% des données sont rejetées, principalement en provenance d'articles traitant de langues exotiques, ou de citations en grec ancien ; 68% du texte des BLOGS luxembourgeois est rejeté comme n'étant pas du luxembourgeois, alors même que nous n'avions retenu que 90 des 400 blogs répertoriés, afin de ne conserver que les blogs contenant effectivement une quantité significative de données en luxembourgeois. Enfin 27% des textes en provenance de la CHAMBRE ont été rejetés : au-delà des passages en français,

Modèles	nb. contextes	nb. heures	WER(%)
amorçage (Ang)	63	200h	79,9
amorçage (Fra)	63	300h	76,1
amorçage (All)	63	52h	68,9
amorçage (pooled)	63	552h	70,7
Non-Sup1	7k	80h	33,6
Non-Sup1	22k	80h	33,2
Non-Sup2	31k	193h	29,3
Non-Sup3	39k	500h	28,4
Non-Sup3, mlp	39k	500h	27,43
Non-Sup4, mlp	50k	1200h	25,6

TABLE 2 – Modèles acoustiques du luxembourgeois

présents dans les débats, ce taux élevé est dû à la présence de rapports écrits en français parmi les transcriptions. Après filtrage, 34 millions de mots ont été conservés à partir des presque 52 millions des textes bruts.

Lexique et modèle de langage Les textes filtrés ont été utilisés pour construire un lexique et des modèles de langage. Pour la liste de mots, les 200k mots les plus probables en provenance ces 7 sources décrites précédemment ont été sélectionnés, de manière à minimiser la perplexité des unigrammes sur un texte de développement. Un taux de mots hors vocabulaire (MHV) 2,35% a ainsi pu être obtenu, significativement plus bas que dans des expériences précédentes (Adda-Decker *et al.*, 2008b). Le dictionnaire de prononciation a été obtenu à partir de cette liste de mots, où les prononciation ont été obtenues à l'aide d'un programme de transformation graphème-phonème, puis corrigées à la main. été En ce qui concerne le modèle de langage, le meilleur 3-gramme obtenu par interpolation des 3-gramme calculés sur les 7 sources obtient un perplexité de 369 sur le jeu de développement.

3.3 Modèles acoustiques et taux d'erreurs

La table 2 résume les données audio utilisées lors des itérations successives pour l'apprentissage non supervisé des modèles acoustiques, avec leur taille respective. La dernière colonne donne le taux d'erreur de mots (*Word Error Rate, WER*) sur le jeu de développement Quaero. La partie supérieure de la table donne les WER obtenus avec les 4 modèles d'amorçage appris sur des données de taille significativement plus importante que celle utilisée utilisée dans la section 2. Bien que le modèle allemand obtienne un taux légèrement plus bas, et par souci de cohérence, nous avons utilisé le modèle *pooled* comme modèle d'amorçage.

Dans les 3 premières itérations, des modèles utilisant des paramètres PLP0 ont été construit, alors que le dernier décodage a utilisé des paramètres MLPPLP0 (les paramètres MLP du système allemand ont été utilisés pour le système luxembourgeois) (Fousek *et al.*, 2011). A chaque itération, la quantité de données audio non-transcrites utilisée pour l'apprentissage a été doublée.

Les modèles sont structurellement similaires à ceux décrits à la section 2. Les modèles de phone

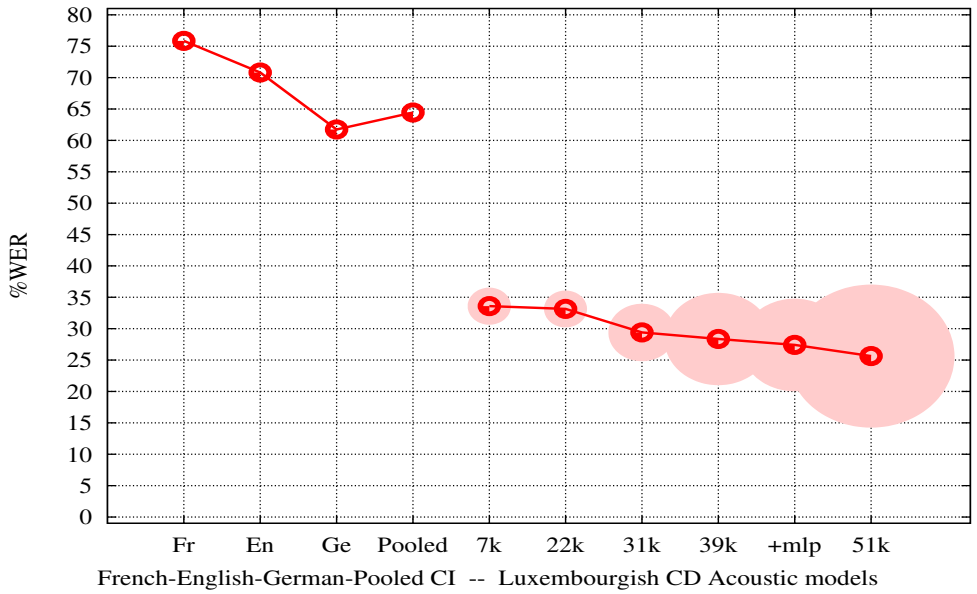

FIGURE 3 – Taux d'erreurs de mots pour les modèles acoustiques non-natifs (gauche) et natifs (droite) la taille du corpus d'apprentissage est représenté par un disque autour du point de résultat.

en contexte triphone sont indépendants du mot, mais dépendants de la position dans le mot et indépendants du genre, étant donné que des expériences précédentes sur l'apprentissage non-supervisé de modèles acoustiques, ont montré que les modèles ainsi obtenus avaient des résultats identiques qu'ils dépendent du genre ou non.

La figure 3 synthétise les taux d'erreurs obtenus. La partie gauche montre les taux pour les modèles non-natifs (anglais, français, allemand, multilingue mutualisé) qui sont compris entre 60% et 75% d'erreurs-mots. La partie droite montre le même type de résultats, mais avec des modèles natifs luxembourgeois obtenus par apprentissage non-supervisé (sélection d'une partie du corpus d'apprentissage luxembourgeois transcrit automatiquement avec les modèles multilingues). Différents jeux de modèles contexte-dépendants luxembourgeois (de 7k à 51k contextes) ont été estimés, le nombre de contextes augmentant progressivement avec la quantité de données sélectionnés pour l'apprentissage. Ce volume d'apprentissage croissant est représenté dans la figure par le diamètre des disques autour du résultat correspondant.

4 Conclusion et perspectives

Le travail présenté ici décrit la construction de modèles acoustiques pour le luxembourgeois, une langue peu dotée ayant subi de fortes influences germaniques et romanes. Nous avons tout d'abord exploré les modèles d'amorçage possibles, et défini un inventaire phonémique que nous avons mis en correspondance avec les inventaires des 3 principales langues exogènes présentes au Luxembourg (allemand, français et anglais). Ensuite , différents modèles d'amorçage ont été construits, monolingues ou multilingues (par concaténation des données d'apprentissage). Ces modèles ont été utilisés lors d'alignements forcés de données audio en luxembourgeois. L'identité du modèle utilisé lors de l'alignement donne des indications sur les correspondances phonémiques

inter-langues, et les modèles acoustiques. Les taux d'utilisation du modèle *pooled* donnent une mesure indicative de l'écart entre les réalisations acoustiques d'une paire donnée de phonèmes dans la langue source et la langue cible. Par exemple, on observe une correspondance importante entre le luxembourgeois et l'anglais pour le phonème /ɛ͡ɪ /. Comme perspective, nous comptons utiliser cette mesure pour des applications d'apprentissage de langues, en mettant en avant pour une paire L1/L2 donnée, la liste des phonèmes potentiellement difficile. Enfin, nous avons utilisé ces modèles d'amorçage pour construire des modèles acoustiques en luxembourgeois à partir d'une très grande quantité de données audio non-transcrites par des méthodes d'apprentissage non-supervisé et un modèle de langage du luxembourgeois appris sur des textes issus du Web et filtrés de manière à éliminer les textes d'une autre langue. Un taux d'erreurs-mots de 25.6% a été obtenu sur le jeu officiel de développement de Quaero.

Références

Adda-Decker, M., Barras, C., Adda, G., Paroubek, P., de Mareïl, P. B. et Habert, B. (2008a). Annotation and analysis of overlapping speech in political interviews. *In LREC*. European Language Resources Association.

Adda-Decker, M., Lamel, L. et Adda, G. (2011a). 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)*, pages 47–50, Poznan, Poland.

Adda-Decker, M., Lamel, L. et Snoeren, N. (2010). 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.

Adda-Decker, M., Pellegrini, T., Bilinski, E. et Adda, G. (2008b). Developments of letzebuergesch resources for automatic speech processing and linguistic studies. *In LREC*.

Adda-Decker, M., Snoeren, N. et Lamel, L. (2011b). Studying Luxembourgish phonetics via multilingual forced alignments. *In ICPhS'11, 17th International Congress of Phonetic Sciences*, Hong Kong, China.

Fousek, P., Lamel, L. et Gauvain, J.-L. (2011). Combining mlp and plp features for speech transcription. *In Handbook of natural language processing and machine translation : DARPA global autonomous language exploitation*, pages 408–416. Springer.

Lamel, L., Gauvain, J. et Adda, G. (2002). Lightly supervised and unsupervised acoustic model training. *Computer Speech and Language*, 16(1):115–229.

Lavergne, T., Cappé, O. et Yvon, F. (2010). Practical very large scale CRFs. *In Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513. Association for Computational Linguistics.

Oparin, I., Lamel, L. et Gauvain, J.-L. (2013). Rapid development of a latvian speech-to-text system. *In ICASSP*, pages 7309–7313. IEEE.

Schanen, F. (2004). *Parlons Luxembourgeois*. L'Harmattan.

Schultz, T. et Waibel, A. (2001). Experiments on cross-language acoustic modeling. *In Proceedings of Eurospeech*, Aalborg.