Alignement de séquences phonétiques pour une analyse phonologique des erreurs de transcription automatique

Camille Dutrey [1,2] Martine Adda-Decker [1,3] Naomi Yamaguchi [1]
(1) Laboratoire de Phonétique et Phonologie (LPP), 19 rue des Bernardins, Paris, France
(2) Laboratoire National de Métrologie et d'Essais (LNE), 29 avenue Roger Hennequin, Trappes, France
(3) Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), Rue John Von Neumann, Orsay, France
camille.dutrey@lne.fr, {madda, naomi.yamaguchi}@univ-paris3.fr

RÉSUMÉ

La transcription automatique de la parole obtient aujourd'hui des performances élevées avec des taux d'erreur qui tombent facilement en dessous de 10% pour une parole journalistique. Cependant, pour des conversations plus libres, ils stagnent souvent autour de 20–30%. En français, une grande partie des erreurs sont dues à des confusions entre homophones n'impliquant pas les niveaux acoustico-phonétique et phonologique. Cependant, de nombreuses erreurs peuvent s'expliquer par des variantes de productions non prévues par le système. Afin de mieux comprendre quels processus phonologiques pourraient expliquer ces variantes spécifiques de la parole spontanée, nous proposons une analyse des erreurs en comparant prononciations attendue (référence) et reconnue (hypothèse) *via* un alignement phonétique par programmation dynamique. Les distances locales entre paires de phonèmes appariés correspondent au nombre de traits phonétiques disjoints. Nos analyses permettent d'identifier les traits phonétiques les plus fréquemment impliqués dans les erreurs et donnent des pistes pour des interprétations phonologiques.

ABSTRACT

Phonetic sequences alignment for a phonemic analysis of automatic speech transcription errors

Nowadays, word error rates of automatic speech transcription systems tend to fall below 10% for journalistic speech. However, in the case of free conversations, error rates remain much higher, typically around 20-30%. Error sources range from system limitations such as out of vocabulary words to speaker production errors. In French, many errors are due to homophonic words, for which neither acoustic-phonetic nor phonological levels are to blame. An important part may be related to production variants unknown to the system. To investigate which phonological processes might contribute to explain fluent speech specific variants, a phone sequence alignment between reference and hypothesis phone strings was implemented using dynamic programming. Local distances are computed as the total number of disagreeing phonetic features between phone pairs. The resulting analyses highlight the features most frequently involved in recognition errors and provide insight for phonological interpretations of fluent speech variation.

MOTS-CLÉS : alignement de séquences, traits distinctifs, programmation dynamique, reconnaissance automatique de la parole, erreurs de transcription.

KEYWORDS: sequence alignment, distinctive features, dynamic programmation, automatic speech recognition, transcription errors.

1 Introduction

Dans cette contribution, nous proposons d'analyser les erreurs de transcription automatique de la parole d'un point de vue phonétique. Les erreurs d'un système de transcription sont habituellement comptabilisées au niveau du mot et une confusion entre deux formes fléchies homophones (p. ex. « politique » et « politiques ») compte autant qu'une erreur entre mots très différents (p. ex. « affaire » et « ferveur » dans la suite « l'affaire Woerth » reconnue comme « la ferveur »).

Pour cela, nous comparons la chaîne phonétique correspondant aux mots de la transcription automatique (hypothèse ou HYP) à celle provenant des mots de la transcription manuelle (référence ou REF). La comparaison est effectuée dans des zones d'erreur, c'est-à-dire aux endroits où le système de transcription produit des mots différents de ceux attendus par la référence. Pour comparer des chaînes de caractère, une mesure fréquemment utilisée est la distance d'édition ou la distance de Levenshtein (Levenshtein, 1965), qui donne le nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre. Dans notre cas, les caractères correspondent à des phones. L'alignement de séquences phonétiques a été utilisé pour de nombreuses recherches en phonologie computationnelle (Kondrak, 2000, 2003), en dialectologie (Heeringa, 2004; Heeringa *et al.*, 2002) ou encore en phonétique clinique (Connolly, 1997).

Nous adoptons ici ce type d'approche pour l'analyse d'erreurs issues de systèmes de transcription automatique de la parole. Nous proposons d'adapter la distance de Levenshtein pour mieux tenir compte de la proximité phonético-phonologique entre phonèmes. Par exemple, le /p/ est plus proche du /b/ que du /s/ ou du /a/.

2 Corpus, approche et méthode

Nos travaux s'inscrivent dans le cadre plus large de recherches menées sur la caractérisation des erreurs produites par des systèmes de transcription de la parole, notamment au sein du projet ANR VERA [1] (Goryainova *et al.*, 2014; Luzzati *et al.*, 2014; Santiago *et al.*, 2015). L'objectif est d'étudier l'impact de ces erreurs sur des applications plus complexes comme l'indexation, la traduction ou le repérage d'entités nommées à partir de flux audio. D'autres finalités consistent à contribuer à une évaluation plus informative des systèmes de transcription et à mieux rendre compte des aspects linguistiques impliqués dans les erreurs. Nous développons ce dernier volet en focalisant sur l'interface phonétique–phonologie dans les erreurs de transcription des sorties du système de reconnaissance du LIUM (Bougares *et al.*, 2013) avec les données de la campagne ETAPE (Gravier *et al.*, 2012). Dans la suite, nous présentons d'abord le corpus qui fournit les erreurs de transcription, c'est-à-dire les séquences phonétiques (REF *vs* HYP) à aligner. Nous développons ensuite le choix de traits phonétiques pour décrire les phonèmes du français, utilisés pour le calcul de distance entre paires de phonèmes. Nous précisons que ce calcul de distance se fait uniquement à partir de traits sans prendre en compte les réalisations acoustiques des sons impliqués. Enfin, nous rappelons brièvement l'algorithme de programmation dynamique tel que mis en œuvre pour notre analyse.

1. http://projet-vera.univ-lemans.fr/.

2.1 Corpus de parole préparée et spontanée : ETAPE

Nous avons travaillé sur un corpus de parole journalistique préparée et spontanée constitué d'émissions radio et télé-diffusées, le corpus ETAPE (Gravier *et al.*, 2012). Nous en avons traité 58 enregistrements, ce qui représente environ 35h de parole pour 339k mots prononcés. La transcription automatique produite par le système du LIUM, comporte 323k mots sur ce sous-ensemble d'ETAPE ; pour une description détaillée de ce système et de sa paramétrisation, se référer à Bougares *et al.* (2013). Afin d'étudier les erreurs de transcription automatique d'un point de vue phonétique, deux étapes préliminaires ont été réalisées :

1. un alignement REF *vs* HYP au niveau des mots à l'aide du NIST Scoring Toolkit [2] pour obtenir les types d'erreur (correct *versus* substitution, suppression et insertion) ;
2. une phonétisation des mots réalisée avec le système d'alignement forcé du LIMSI (Gauvain *et al.*, 2003) et un jeu de 33 phonèmes du français : à noter l'absence du /œ/ supplanté par /ə/. L'alignement est réalisé sur l'ensemble de l'audio, à la fois pour la référence et l'hypothèse, en utilisant le même dictionnaire de prononciation (dictionnaire du LIUM) que celui utilisé par le système de reconnaissance de la parole.

REF	« donc	le	**fort**	**taux**	**de**	natalité »
HYP	« donc	le	*	*	**forte**	natalité »
erreur	C	C	D	D	S	C

FIGURE 1 – Extrait de parole transcrite avec alignement REF *versus* HYP et indication du type d'erreur assignée par le système pour chaque mot (C = correct ; D = suppression ; S = substitution).

Nous avons extrait du corpus 18 051 zones d'erreurs : une zone d'erreur correspond à une suite ininterrompue de mots erronés entre deux mots non erronés. La figure 1 donne un exemple d'énoncé où la zone d'erreur est marquée en gras. Les données ont été pré-traitées de manière à en exclure les zones d'erreur trop complexes résultant souvent de décalages entre REF et HYP. Cette sélection, qui exclut 13,6 % des zones d'erreur, s'appuie sur des critères liés à la différence de longueur entre la séquence phonétique de REF et celle de HYP. Nous avons également choisi d'écarter les zones d'erreur incluant des phénomènes particuliers de la parole spontanée, qui seront analysés à part : hésitations vocaliques, présence d'amorces de mot, *etc.* Au final, 16,5 % des zones d'erreur ont ainsi été mises de côté et 13 021 zones d'erreurs sont conservées.

2.2 Spécification en traits des phonèmes du français

La comparaison des phonèmes du français s'appuie sur la construction d'une matrice de traits distinctifs : les traits pris en compte sont uniquement les traits distinctifs en français et sont adaptés de Sagey (1986) et Walker (1993). Ces 13 traits ont été considérés comme privatifs (Mester & Ito, 1989) pour la spécification des phonèmes du français. La spécification est par ailleurs totale : tous les phonèmes sont spécifiés pour tous les traits, y compris lorsque la valeur de trait est redondante (p. ex. [voisé] pour les sonantes), comme présenté dans le tableau 1. Dans cette perspective, nous parlerons alors de traits phonétiques. Dans cette étude, le terme « phonème » englobe consonnes, voyelles et semi-voyelles, même si ces dernières peuvent être considérées comme des allophones de certaines voyelles.

2. Outil accessible à l'adresse Web suivante : http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sctk.htm.

/p/	cons., labial
/b/	cons., labial, voisé
/t/	cons., coronal
/d/	cons., coronal, voisé
/k/	cons., dorsal
/g/	cons., dorsal, voisé
/f/	cons., cont., labial
/v/	cons., cont., labial, voisé
/s/	cons., cont., coronal
/z/	cons., cont., coronal, voisé
/ʃ/	cons., cont., coronal, post.
/ʒ/	cons., cont., coronal, post., voisé
/l/	cons., cont., coronal, sonant, voisé, latéral
/m/	cons., cont., labial, sonant, voisé, nasal
/n/	cons., cont., coronal, sonant, voisé, nasal
/ɲ/	cons., cont., coronal, sonant, voisé, nasal, post.
/ʁ/	cons., cont., dorsal, sonant, voisé

(a) Consonnes.

/i/	cont., coronal, sonant, voisé, haut
/y/	cont., coronal, sonant, voisé, haut, arrondi
/u/	cont., dorsal, sonant, voisé, haut, arrondi
/e/	cont., coronal, sonant, voisé
/ø/	cont., coronal, sonant, voisé, arrondi
/o/	cont., dorsal, sonant, voisé, arrondi
/ɛ/	cont., coronal, sonant, voisé, bas
/ə/	cont., sonant, voisé
/ɔ/	cont., dorsal, sonant, voisé, bas, arrondi
/a/	cont., dorsal, sonant, voisé, bas
/ɔ̃/	cont., dorsal, sonant, voisé, bas, arrondi, nasal
/ɛ̃/	cont., sonant, voisé, bas, nasal
/ɑ̃/	cont., dorsal, sonant, voisé, bas, nasal
/j/	cont., coronal, sonant, voisé, haut
/w/	cont., dorsal, sonant, voisé, haut, arrondi
/ɥ/	cont., coronal, sonant, voisé, haut, arrondi

(b) Voyelles et semi-voyelles.

TABLE 1 – Spécification en traits pour les phonèmes du français utilisée pour le calcul de distance et l'alignement de séquences phonétiques (cons. = consonantique ; cont. = continu ; post. = postérieur).

Le trait [consonantique] distingue les consonnes des voyelles et semi-voyelles, et représente le degré de constriction dans le conduit vocal. Les traits de lieu [labial], [coronal] et [dorsal] indiquent le lieu d'articulation des sons ; le trait [postérieur] distingue dans les consonnes [coronal] les sons produits avec l'arrière de la langue de ceux produits avec l'avant de la langue. Le trait [voisé] désigne le voisement des phonèmes. Le trait de mode [sonant] distingue les sonantes des obstruantes ; [continu] indique le passage continu de l'air dans le conduit vocal, et distingue les sonantes et fricatives des occlusives. [nasal] est spécifié pour les sons laissant passer l'air par la cavité nasale. Les traits vocaliques [haut] et [bas] caractérisent l'aperture. Le trait [arrondi] spécifie les (semi-)voyelles produites avec un arrondissement des lèvres. Chaque phonème est ainsi représenté par un vecteur de dimension 13 (V_φ) avec des 0 pour tous les traits non-spécifiés et des 1 pour les traits spécifiés.

2.3 Calcul de distances pour la comparaison de paires de phonèmes

Nous avons calculé, pour chaque paire de phonèmes (φ_i, φ_j) du français [3], une distance phonétique $d(i, j)$ (*cf.* section 2.4) à partir de la spécification en traits présentée ci-dessus. La distance $d(i, j)$ correspond à la somme de l'opérateur ou-exclusif sur les 13 dimensions des vecteurs correspondant aux phonèmes φ_i et φ_j ($V_{\varphi_i} \oplus V_{\varphi_j}$) et donne le nombre de traits phonétiques disjoints. Alors que théoriquement la valeur maximale pourrait être 13 pour un vecteur de dimension 13, il se trouve que les traits sont distribués de telle manière que la distance maximale se limite à 9. La distribution de ces distances est présentée en figure 2, selon le type de paire : voyelle *versus* voyelle ; consonne *versus* consonne ; voyelle *versus* consonne. 36 paires ont une distance nulle : il s'agit de comparaisons de phonèmes identiques, en tenant compte des allophones. Ainsi, les paires /i/-/j/, /y/-/ɥ/ et /u/-/w/ sont également distantes de 0.

3. Soit 561 combinaisons ; le nombre de paires correspond au nombre de cases d'une matrice triangulaire hors diagonale ($33 \times 32/2$) plus la diagonale (33).

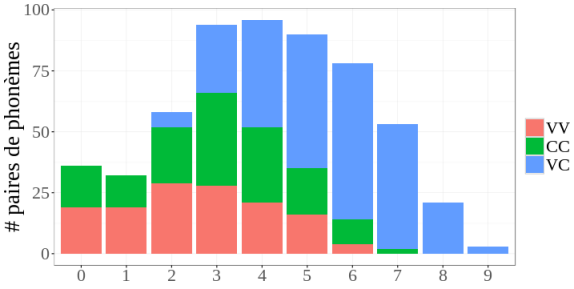


FIGURE 2 – Distribution des paires de phonèmes en fonction de leur distance locale, avec caractérisation du type de paire comparée (VV = voyelles ; CC = consonnes ; VC = voyelle *vs* consonne).

On peut remarquer que les distances faibles correspondent majoritairement à des paires "homogènes" VV et CC et que la proportion de paires CV augmente au fur et à mesure que la distance augmente.

Le tableau 2 recense les paires de phonèmes appartenant aux distances maximales et minimales (hors distances nulles) pour chaque type de paires (VV, CC et VC). On remarque la présence de /p/ et /t/ dans les paires de CV maximalement distantes : ceci peut s'expliquer par le fait que ces phonèmes, particulièrement /t/, sont considérés comme des sons phonologiquement non marqués (Paradis & Prunet, 1991) et sont de ce fait spécifiés par moins de traits que les autres phonèmes. Les distances entre phonèmes sont utilisées dans cette étude comme connaissance linguistique dans le programme d'alignement de séquences phonétiques par programmation dynamique.

Voyelle-Voyelle	min=1	/i-y/ /i-e/ /i-ɥ/ /y-ø/ /y-j/ /u-o/ /e-ø/ /e-ɛ/ /e-ə/ /e-j/ /ø-ɥ/ /o-ɔ/ /o-w/ /ɔ-a/ /ɔ̃-ɔ̃/ /a-ɑ̃/ /ɔ̃-ɑ̃/ /ɛ̃-ɑ̃/ /j-ɥ/
	max=6	/i-ɔ̃/ /y-ɑ̃/ /ɔ̃-j/ /ɑ̃-ɥ/
Consonne-Consonne	min=1	/p-b/ /p-f/ /b-v/ /t-d/ /t-s/ /d-z/ /k-g/ /f-v/ /s-z/ /s-ʃ/ /s-/ /ʃ-ʒ/ /n-ɲ/
	max=7	/p-ɲ/ /k-ɲ/
Voyelle-Consonne	min=2	/e-z/ /e-l/ /e-n/ /o-ʁ/ /ə-ʁ/ /a-ʁ/
	max=9	/ɔ̃-p/ /ɔ̃-t/ /ɔ̃-ʃ/

TABLE 2 – Paires de phonèmes impliquées dans les distances minimales et maximales par type (VV = paires de voyelles ; CC = paires de consonnes ; VC = paires de voyelle *versus* consonne).

2.4 Mesure de distances phonétiques et alignement de séquences

Le programme d'alignement de séquences phonétiques est adapté de l'algorithme de Levenshtein (1965) qui permet de calculer des mesures de distances entre deux chaînes de caractères. Il s'appuie sur la programmation dynamique (Bellman, 1957; Vintsyuk, 1968), dont le principe est rapidement décrit ci-dessous. Soient deux séquences phonétiques $\Phi_I = \varphi_1\varphi_2 \dots \varphi_I$ (hypothèse) et $\Phi_J = \varphi_1\varphi_2 \dots \varphi_J$ (référence) de longueur I et J respectivement. On impose le départ de l'alignement au début des deux chaînes respectives ce qui se traduit par des conditions d'initialisation $D(0,0) = 0$, $D(0,j) = \infty$ pour $j > 0$, $D(i,0) = \infty$ pour $i > 0$. Ensuite la récurrence sur i, j (chaînes partielles de 1 à i et de 1 à j) s'écrit comme suit :

$$D(i,j) = min \begin{cases} D(i-1,j) + d(i,j) & \text{insertion} \\ D(i,j-1) + d(i,j) & \text{omission} \\ D(i-1,j-1) + 2 \times d(i,j) & \text{correct ou substitution} \end{cases} \quad (1)$$

où $d(i,j) = 0$ si $\varphi_i = \varphi_j$. L'arrêt se fait naturellement à (I, J), à la fin des deux chaînes. Pour récupérer l'alignement correspondant à la distance globale minimale, nous avons introduit dans la récurrence une matrice de *retour-arrière*, qui à chaque point (i,j) garde la mémoire du meilleur point précédent (argmin de l'équation 1). La distance globale, qui permet de caractériser la dissimilarité phonétique entre deux séquences de phonèmes, est ensuite normalisée par le nombre de phonèmes dans la référence. Elle nous permet d'analyser plus finement les erreurs commises par les systèmes de transcription de la parole, notamment grâce à son appui sur des connaissances phonologiques.

Dans un premier temps, nous souhaitons limiter nos analyses à des zones d'erreurs que nous jugeons intéressantes d'un point de vue phonétique comme phonologique. Dans ce but, nous rejetons dans la suite celles dont les longueurs sont très différentes entre REF et HYP et pour lesquelles des problèmes de découpage du signal ou de bruit de fond viennent supplanter les facteurs linguistiques. Nous gardons ainsi 11 753 zones d'erreur des données ETAPE, dont les distances globales normalisées se distribuent entre 0 et 5 (*cf.* figure 4). Pour ces zones d'erreur, l'information de *retour-arrière* a été utilisée pour récupérer l'alignement de la séquence phonème à phonème. En effet, l'alignement des zones d'erreur au niveau du phonème, comme illustré en figure 3, permet de mieux décrire et analyser les erreurs produites par le système de transcription de la parole d'un point de vue phonétique. Comme le met en exergue l'exemple illustré, cet alignement permet d'obtenir une finesse de localisation et d'identification phonétique des erreurs totalement absente des méthodes classiques qui évaluent les systèmes de reconnaissance automatique de la parole au seul niveau de la séquence de mots.

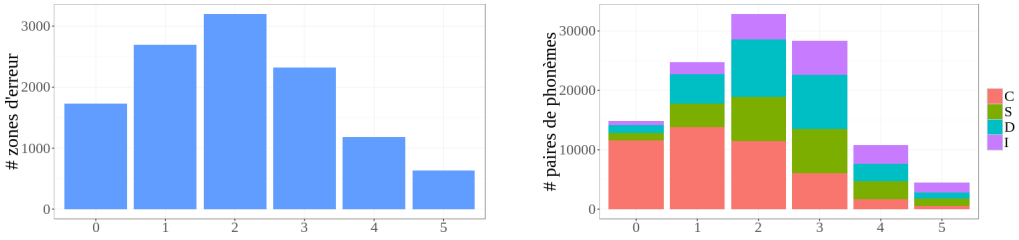
REF	**fort**	**taux**	**de**	⇒	[	f	ɔ	ʁ	t	**o**	**d**	]
HYP			**forte**	⇒	[	f	ɔ	ʁ	t	**ə**	*****	]
erreur	D	D	S			C	C	C	C	S	D	

FIGURE 3 – Extrait de parole transcrite avec comparaison de l'alignement REF *versus* HYP produit par un système d'évaluation de la transcription automatique (au niveau du mot) et de celui produit par le système d'alignement de séquences phonétiques (C = correct ; D = suppression ; S = substitution).

La figure 3 illustre une zone d'erreur de distance globale normalisée égale à 1 pour laquelle l'évaluation classique produit deux omissions de mots et une substitution et qui, d'un point de vue phonétique, se révèle majoritairement correcte. L'erreur commise peut s'expliquer d'un côté par une forte réduction temporelle de l'article « de » dans le contexte « fort taux de natalité » : le schwa est tombé et le [d] se limite à environ 30 ms de barre de voisement avant le [n]. Dans ce contexte de quasi-absence du *de*, le modèle de langage impose « forte natalité » plutôt que « fort taux de natalité ».

3 Analyse des erreurs de transcription automatique

La méthode présentée dans cette étude, permettant de calculer des distances entre chaînes phonétiques au-delà des mots et d'aligner ces dernières en tenant compte d'informations phonologiques, peut être mise au service d'une analyse linguistique des erreurs de transcription de la parole. Nous souhaitons ainsi contribuer à l'évaluation des systèmes de transcription de la parole en utilisant ces informations de manière à mieux identifier les variations impactant les phonèmes et le rôle des traits phonétiques.



(a) distribution des zones d'erreur.

(b) distribution des paires de phonèmes.

FIGURE 4 – Caractérisation des zones d'erreur : (a) distribution des zones d'erreur en fonction de la distance globale normalisée entre REF et HYP ; (b) distribution des paires de phonèmes impliquées par type d'erreur subie en fonction de la distance globale normalisée (C = correct ; S = substitution ; D = suppression ; I = insertion).

3.1 Caractérisation de zones d'erreurs par distance phonétique

La figure 4 permet de visualiser la distribution des zones d'erreur en fonction de leur distance phonétique. Une analyse préliminaire qualitative permet de faire l'hypothèse que cette mesure de distance pourrait permettre de catégoriser efficacement les zones d'erreur produites par les systèmes de transcription automatique. En effet, les zones présentant une distance normalisée nulle permettent d'identifier des chaînes homophoniques ou quasi-homophoniques, telles « leaders » *versus* « lits de leur » ou « base » *versus* « basse », y compris sur des séquences relativement longues, comme « vin de Féternes » *versus* « vingt-deux faits termes ».

Les séquences présentant une distance moyenne (p. ex. 2, distance la plus fréquente du corpus) sont phonétiquement proches tout en présentant, notamment, de nombreuses substitutions sur des paires de type VV ou CC, comme dans « que ce label » [ksəlabɛl] *versus* « solennel » [sɔlanɛl] ou des prononciations non-canoniques sources de confusion, comme dans « sans sans langue de bois » [sɑ̃sɑ̃lɑ̃gədəbwa] *versus* « cinq cent emplois » [sɛ̃ksɑ̃ɑ̃plwa] (pour cet exemple, la transcription automatique est également mise en difficulté par de la parole superposée). Enfin, les séquences présentant une distance maximale (p. ex. « bon Copé » *vs* « à la rentrée ») sont souvent particulièrement difficiles à transcrire, avec beaucoup de parole superposée ou de bruit environnant.

3.2 Implication des traits phonétiques dans les zones d'erreurs

L'examen des traits phonétiques impliqués dans les zones d'erreur (*cf.* figure 5) indique que tous les traits n'ont pas la même importance dans les erreurs. Les traits les mieux reconnus sont les traits [continu], [voisé], [sonant], [consonantique] qui sont les traits les plus fréquents, et qui correspondent à des distinctions fondamentales phonologiques (Clements, 1985). Les traits qui sont plus souvent substitués que bien reconnus sont les traits [arrondi] et [postérieur], qui ne distinguent respectivement qu'une petite partie des voyelles et des consonnes. Quant aux suppressions et aux insertions, elles concernent majoritairement les traits qui sont partagés par de nombreux phonèmes : [continu] et [voisé] (spécifiés pour 27 phonèmes), [sonant] (spécifié pour 21 phonèmes). Ces premières observations semblent indiquer que les traits participent dans les différents types d'erreurs en fonction de leur rôle dans le système phonologique et de leur place dans une séquence de sons. Ces résultats sont bien

entendu à approfondir par l'analyse des combinaisons de traits impliqués dans chaque type d'erreur, et par l'étude des séquences de traits dans les suites de phonèmes.

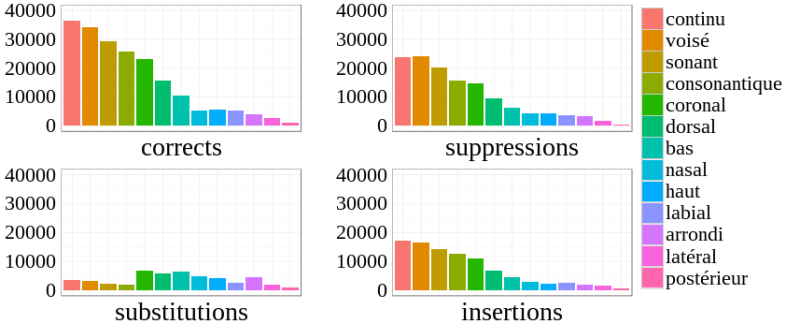


FIGURE 5 – Distribution des traits phonétiques selon leur implication dans des types d'erreurs.

4 Conclusion

Nous avons présenté nos travaux sur l'analyse des erreurs produites d'un système de transcription de l'évaluation ETAPE. Afin d'analyser des zones d'erreur (incluant tous les mots erronés entre les deux extrémités bien reconnues), nous avons introduit une nouvelle méthodologie s'appuyant sur la programmation dynamique en traitement automatique et les traits phonétiques. Cette approche vise à aligner non plus les mots en tant que tels, mais leurs séquences phonétiques respectives afin de mieux décrire les erreurs en matière de proximité phonique. L'utilisation des traits met en lumière le rôle de certains traits, eux-mêmes importants dans le système phonologique, dans les erreurs du système de transcription. En particulier, ce résultat apporte des arguments en faveur de la structuration du système phonologique en traits hiérarchisés (Clements, 1985). Les résultats permettent de trier les zones d'erreur suivant une distance phonétique : à distance faible, nous sommes en présence d'erreurs homophones et quasi-homophones. Au sein de ce sous-ensemble, il sera intéressant d'étudier plus finement les processus phonétiques et phonologiques (lénition, assimilation, chute de segments, réductions et divers metaplasmes). Plusieurs améliorations de la procédure d'alignement sont prévues, et notamment : affiner la représentation des traits phonétiques (p. ex. type de spécification) ; améliorer le calcul des distances locales en tenant compte du voisinage phonétique immédiat. Nous avons également pour perspective de développer les analyses à l'interface phonétique-phonologie, de produire des analyses d'erreur en contexte (analyse des triphones) et fonction des frontières de mots/syllabes et de mieux rendre compte des phénomènes de réduction en parole spontanée. Enfin, nous envisageons dans le futur d'ajouter un décodage phonétique afin d'étudier le rôle des traits phonétiques dans les erreurs de reconnaissance automatique hors contraintes lexicales (et hors contraintes de plus haut niveau).

Remerciements

Ce travail a été partiellement financé par l'Agence Nationale de la Recherche au titre du projet VERA (ANR-12-BS02-006-01) et du programme Investissements d'Avenir (ANR-10-LABX-0083).

Références

BELLMAN R. (1957). *Dynamic Programming*. Princeton University Press.

BOUGARES F., DELÉGLISE P., ESTÈVE Y. & ROUVIER M. (2013). LIUM ASR system for ETAPE French evaluation campaign : experiments on system combination using open-source recognizers. In *6th International Conference on Text, Speech and Dialogue (TSD'13)*.

CLEMENTS G. N. (1985). The geometry of phonological features. *Phonology*, **2**, pp. 225–252.

CONNOLLY J. (1997). Quantifying target-realization differences. Part I : Segments. *Clinical Linguistics & Phonetics*, **11**, pp. 267–287.

GAUVAIN J., LAMEL L., SCHWENK H., ADDA G., CHEN L. & LEFÈVRE F. (2003). Conversational telephone speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*.

GORYAINOVA M., GROUIN C., ROSSET S. & VASILESCU I. (2014). Morpho-Syntactic Study of Errors from Speech Recognition System. In *LREC'14*, p. 3050–3056.

GRAVIER G., ADDA G., PAULSSON N., CARRÉ M., GIRAUDEL A. & GALIBERT O. (2012). The ETAPE Corpus for the Evaluation of Speech-based TV Content Processing in the French Language. In *8th International Conference on Language Resources and Evaluation (LREC'12)*.

HEERINGA W. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. PhD dissertation, Rijksuniversiteit Groningen, Groningen.

HEERINGA W., NERBONNE J. & KLEIWEG P. (2002). Validating Dialect Comparison Methods. In *24th Annual Meeting of the Gesellschaft für Klassifikation (GFKL'02)*, p. 445–452 : Springer.

KONDRAK G. (2000). A New Algorithm for the Alignment of Phonetic Sequences. In *6th Applied Natural Language Processing Conference (ANLP'00)*, p. 288–295.

KONDRAK G. (2003). Phonetic Alignment and Similarity. *Computers and the Humanities*, **37 (3)**, pp. 273–291.

LEVENSHTEIN V. I. (1965). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Doklady Akademii Nauk SSSR*, **163**, pp. 845–848.

LUZZATI D., GROUIN C., VASILESCU I., ADDA-DECKER M., BILINSKI E., CAMELIN N., KAHN J., LAILLER C., LAMEL L. & ROSSET S. (2014). Human annotation of ASR error regions : Is "gravity" a sharable concept for human annotators ? In *9þth International Conference on Language Resources and Evaluation (LREC'14)*, p. 3050–3056.

MESTER R. A. & ITO J. (1989). Feature Predictability and Underspecification : Palatal Prosody in Japanese Mimetics. *Language*, **65 (2)**, pp. 258–293.

PARADIS C. & PRUNET J.-F. (1991). Introduction : Asymmetry and Visibility in Consonant Articulations. In C. PARADIS & J.-F. PRUNET, Eds., *The Special Status of Coronals : Internal and External Evidence*, p. 1–28. San Diego : Academic Press.

SAGEY E. C. (1986). *The Representation of Features and Relations in Non-linear Phonology*. PhD thesis, Massachusetts Institute of Technology.

SANTIAGO F., DUTREY C. & ADDA-DECKER M. (2015). Towards a Typology of ASR Errors via Syntax-Prosody Mapping. In *Errors by Humans and Machines in multimedia, multimodal and multilingual data processing (ERRARE'15)*.

VINTSYUK T. (1968). Speech Discrimination by Dynamic Programming. *Kibernetika*, **4**, pp. 81–88.

WALKER R. (1993). A Vowel Feature Hierarchy of Contrastive Specification. *Toronto Working Papers in Linguistics*, **12 (2)**, pp. 179–198.