Aspects relatifs à la transcription phonétique interactive du signal audio[1]

LUMINIŢA BOTOŞINEANU[2], ELENA MUSCĂ[3],
FLORIN-TEODOR OLARIU[2], IOAN PĂVĂLOI[3]

The paper presents the results obtained in the interactive phonetic transcription for the Romanian vowels [a], [e] and [i]. After a brief introduction, the resource creation process is presented. A subset of video recordings are selected from the site of *Atlasul lingvistic audiovizual al Bucovinei* (ALAB) [Audiovisual Linguistic Atlas of Bucovina]. The audio parts, extracted and manually annotated using the Praat software, were phonetically transcribed using the ALR_IIT editor. These result in six sets of feature vectors generated using F0-F3 formant values, MFCC (Mel-Frequency Cepstral Coefficients) coefficients and PLP (Perceptual Linear Prediction) coefficients. For the recognition we used two discriminative classification algorithm: k-NN, for k=1, k=3 and k=5, and SVM. For the k-NN algorithm, we used three distances: Euclidian, Manhattan and Canberra. The results obtained for [a], [e] and [i] are then discussed in detail. The last section is dedicated to the conclusions and to the prospects of our research.

Key-words: *interactive phonetic transcription*, *ALR_IIT*, *MFCC*, *PLP*, *k-NN*, *SVM*

Introduction

A travers plus de 100 ans de recherches roumaines en cartographie linguistique, on a publié pas moins de trois séries d'atlas

[1] Cet article a été rédigé dans le cadre du projet „Atlasul lingvistic audiovizual al Bucovinei (ALAB) – faza a II-a", financé par UEFISCDI (code du projet: PN-II-RU-TE-2014-4-0880).

[2] Institut de Philologie Roumaine „A. Philippide", Iaşi, Roumanie.
[3] Institut d'Informatique Théorique, Iaşi, Roumanie.

linguistiques nationaux, ainsi que de nombreux travaux cartographiques dédiés à des aires ethnolinguistiques particulières, situées au nord ou au sud du Danube. Cette riche expérience en domaine est devenue, à partir de 2000, un solide point de départ vers la modernisation et la mise à jour des méthodes de travail, principalement en utilisant la technologie informatique dans l'édition des atlas linguistiques. L'un des premiers résultats notables à cet égard a été enregistré en 2007 dans le centre universitaire de Iaşi, où on a publié le premier volume de la série *Le Nouvel Atlas linguistique roumain, par régions* (NALR) édité entièrement à l'aide de l'ordinateur, à savoir le volume III de NALR. *Moldavie et Bucovine*. L'un des principaux avantages de l'édition informatisée des atlas, qui a été compris par les chercheurs depuis le début de cet effort de modernisation de la géolinguistique, c'est que la réalisation des atlas linguistiques en format électronique permettra par la suite que le grand volume de données stockées en format numérique puisse être utilisé pour la mise en œuvre de certains futurs projets interdisciplinaires, se proposant, par exemple, une corrélation entre la transcription phonétique et le segment correspondant du signal audio acquis.

Dans le même centre universitaire de Iaşi on a lancé à partir de 2010 un nouveau projet de géolinguistique, qui, par une approche interdisciplinaire (continuant les collaborations antérieures dans le domaine de l'informatisation de la cartographie linguistique roumaine), s'est proposé de réaliser le premier atlas multimédia dédié à la langue roumaine: *Atlas linguistique audiovisuel de la Bucovine* (ALAB). Dans la mise en place de ses perspectives méthodologiques, cet atlas a tiré profit à la fois de la tradition roumaine en domaine et des expériences européennes les plus récentes dans l'édition des atlas linguistiques (*Sprachatlas des Dolomitenladinischen und angrenzender Dialekte* – ALD, *Atlas linguistique audiovisuel du Valais romand* – ALAVAL, etc.).

La transcription phonétique interactive du signal audio

La transcription phonétique du matériel linguistique inclus dans les atlas est actuellement une opération effectuée uniquement par un opérateur humain, à l'aide d'un éditeur spécialement conçu dans ce

but (Apopei et al. 2003, Bejinariu et al. 2006). À l'avenir, cette opération pourrait être faite d'une manière interactive, l'ordinateur proposant, à partir du signal audio, une ou plusieurs variantes de transcription phonétique pour un mot, dont le linguiste puisse choisir la variante optime pour le segment audio envisagé. De la sorte, si l'objectif initial des recherches était celui de parvenir à un instrument utile pour l'édition et la publication des atlas linguistiques, dans cette nouvelle phase, l'accent sera mis sur la création d'un outil propre à soutenir l'opérateur humain dans la transcription phonétique du matériel audio. L'utilité de ce processus est motivée par le fait que la transcription phonétique est une opération chronophage et, par ailleurs, assez difficile. Dans cet article, nous avons l'intention de montrer les premières étapes qu'on a parcourues jusqu'à présent pour concevoir et mettre en œuvre un logiciel pour la transcription phonétique semi-automatique des enregistrements audio dialectaux.

La création des ressources acoustiques et phonétiques

Pour un nombre important de langues (dont le roumain), les ressources acoustiques (bases de données d'enregistrement audio) sont insuffisantes, donc la conception et la création d'un système de transcription phonétique automatique doivent comprendre tout d'abord une activité considérable de collecte de ces ressources.

Pour pouvoir opérer avec des données convenables tant de point de vue qualitatif que quantitatif, on a pris comme point de départ les enregistrements vidéo faites au cours d'un projet récent de cartographie linguistique roumaine, à savoir l'*Atlas linguistique audiovisuel de la Bucovine* (ALAB) (http://www.philippide.ro/alab/). Cette base de données comprend actuellement environ 3 500 réponses en roumain (plus environ 500 réponses houtzoules, enregistrées dans la localité Brodina), ansi qu'une série d'enregistrements spécifiques (socio- et ethnotextes), obtenus des 28 informateurs interrogés au cours de la première phase du projet.

La réalisation de l'archive dialectale

Les enquêtes de terrain à partir desquelles a été créée l'archive ALAB ont été menées en 2011–2013 par une équipe de dialectologues de l'Institut de Philologie Roumaine « A. Philippide » de l'Académie Roumaine, Filiale de Iaşi, en utilisant un set de 126 questions tirées du questionnaire du *Nouvel Atlas linguistique roumain* (le chapitre *La Cour*). Les fichiers vidéo obtenus par la suite du traitement et de l'annotation des enregistrements réalisés ont constitué la collection des données qui a été le point de départ pour l'élaboration d'un site à travers duquel les résultats de l'enquête ont été mis à la disposition de tous les spécialistes intéressés.

Excepté l'examen de la *variation diatopique* (en tant que sujet préféré des travaux de géographie linguistique), dans la réalisation de l'*Atlas linguistique audiovisuel de la Bucovine* on a envisagé la documentation de la variation linguistique à deux niveaux, dans chaque localité étudiée : la variation diagénérationnelle et la variation diasexuelle. Suite à ces options méthodologiques, on a établi que, pour chaque localité, on va enquêter quatre sujets, deux sujets âgés (un homme et une femme ayant plus de 60 ans) et deux jeunes (toujours un homme et une femme, les deux ayant moins de 35 ans). En général on a réussi à respecter ces limites d'âge, mais il y avait aussi des cas où (à cause de la difficulté de trouver des informateurs jeunes dans les zones rurales) on a accepté aussi des sujets légèrement plus âgés par rapport à la limite de 35 ans établie initialement[4].

Pour chaque sujet on a réalisé tout d'abord une interview du type semi-directif, dont le but était surtout d'obtenir certains ethnotextes sur les principales fêtes de l'année (religieuses ou relevant de la sphère de la vie sociale – mariage, baptême, etc.), le choix de ces sujets étant motivé principalement par leur potentiel narratif. Se proposant d'aider les sujets à surmonter le trac initial, amplifié par la présence du caméscope, ce type d'interviews ont précédé l'enquête proprement-dite. On a envisagé une limitation de

[4] On peut trouver des données supplémentaires relatives aux sujets de l'enquête à l'adresse : http://www.philippide.ro/alab_old/inf.php.

la durée de l'enquête à environ 90 (ou bien tout au plus 120) minutes, la disponibilité du sujet étudié se réduisant avec le temps.

Puisqu'il s'agit d'un projet se proposant d'enregistrer et de mettre en évidence le spécifique linguistique des communautés étudiées, en termes de statut social des informateurs, on a opté pour la typologie classique des sujets étudiés, telle qu'elle a été établie par les travaux de pionnier dans le domaine de la dialectologie (*Atlas linguistique de la France*, *Atlas linguistique roumain*, etc.). Ainsi, on a choisi des sujets ayant vécu aussi longtemps que possible dans la communauté-cible (au cours de la scolarité, ainsi qu'après cela).

Les enregistrements ont été réalisés en sept points d'enquête : Ilişeşti, Doroteia, Mănăstirea Humorului, Solca, Deluţ, Fundu Moldovei, Brodina, se retrouvant aussi dans le réseau du *Nouvel Atlas linguistique roumain, par régions*. *Moldavie et Bucovine*.

Les réponses audio-vidéo ainsi obtenues peuvent se limiter à un noyau, représenté par l'équivalent dialectal des *realia* visées par le questionnaire, mais des fois elles sont accompagnées par certaines explications supplémentaires, le sujet faisant des précisions métalinguistiques ou bien des descriptions de contenu en référence aux réalités couvertes par la question.

On a établi un système de codage des réponses obtenues au cours des enquêtes, pour qu'on puisse les enregistrer dans la base de données du projet. La formule générale requise pour le codage des réponses a été construite de sorte qu'elle prenne en compte les principaux paramètres de variation utilisés dans les enquêtes, à savoir :

a) localité;
b) terme / notion documentée;
c) sexe;
d) âge.

On a codé donc le nom des sept localités étudiées (Ilişeşti = il; Mănăstirea Humorului = mh; Solca = so; Brodina = br; Doroteia = do; Deluţ = de; Poiana Stampei = ps). À chaque question on a attribué un code numérique unique composé de trois chiffres (001–999), correspondant au numéro d'ordre dans le questionnaire général. Pour ce qui est du sexe du sujet, on a utilisé de un seul

caractère, [f] = ‘femme’ [b] = ‘mâle’ et pour le groupe d’âge on a choisi [t] = ‘jeune’ et [v] = ‘âgé’. Par conséquent, dans le codage, les deux premiers caractères permettent l’identification du point d’enquête, les trois chiffres qui suivent indiquent la question (de 817 à 987), tandis que les deux caractères suivants assurent l’identification du sujet qui a fourni cette réponse. S’il y a un caractère [n] à la fin du codage, celui-ci indique qu’il s’agit d’une « note » supplémentaire où l’informateur fait certains commentaires marginaux. Ainsi, les fichiers vidéo inclus dans la base de données peuvent être facilement identifiés: par exemple, le code [mh923ft] montre que la séquence vidéo envisagée correspond à la question 923 du questionnaire et la réponse a été offerte par la jeune femme du point d’enquête Mănăstirea Humorului.

La sélection d’un sous-ensemble de données et son traitement

Sur le total de 4 233 enregistrements audio (obtenus à partir des enregistrements vidéo du site ALAB: les réponses proprement-dites, ainsi que les notes supplémentaires), on a effectué la sélection d’un sous-ensemble de plus de 700 items. Pour choisir de la quantité totale de réponses un ensemble de mots d’intérêt (pour effectuer par la suite le traitement acoustique et la transcription phonétique automatique des voyelles [a], [e] et [i] extraites du signal audio), on a mis en place le programme VowelSelect. Celui-ci lit un fichier texte contenant les réponses au questionnaire. Le résultat du roulement du programme est un fichier texte qui comprend les mots contenant les trois voyelles à la fois, ainsi que des mots présentant deux sur les trois voyelles visées ou bien juste une seule. Nous avons opéré ce choix afin d’obtenir une vision globale du matériel enregistré, mais aussi dans la perspective des éventuelles sélections futures.

Ainsi, à partir des réponses au questionnaire et prenant en compte seulement l’aspect graphique, on a obtenu la typologie suivante: a) trois mots contenant les trois graphèmes *a*, *e* et *i* : *sanie* ‘traîneau’, *cireadă* ‘troupeau de vaches’ et *oaie* ‘brebis’; b) 11 mots contenant les graphèmes *a* et *e* : *cea* ‘exclamation pour faire tourner les boeufs attelés à droite’, *ţesală* ‘étrille’, *lapte* ‘lait’, *(vacă) stearpă* ‘(vache) stérile’, *(calul) nechează* ‘(le cheval) hennit’, *găleată* ‘seille, seau’, *strecurătoare* ‘passoire’, *purcea*

'porcelet (femelle)', *castrez* '(je) castre', *căţea* 'chienne', *creastă* 'crête'; c) 12 mots contenant les graphèmes *a* et *i* : *inima (carului)* 'pièce en bois qui relie l'avant à l'arrière du char', *izlaz* 'pacage communal', *iapă* 'jument', *râncaci* 'étalon monorchide', *mioară* 'agnelle', *mia* 'agnelle', *iadă* 'chevrette', *cioban* 'berger', *baci* 'maître-berger, fromager', *(pisica) miaună* '(le chat) miaule', *cuibar* 'nid (de poule), pondoir', *aripi* 'ailes'; d) 13 mots contenant les graphèmes *e* et *i* : *osie* 'essieu', *oişte* 'timon (du char)', *iesle* 'crèche', *putinei* 'baratte', *viţel* 'veau', *vite* 'gros bétail', *herghelie* 'troupeau de chevaux', *împiedic (calul)* 'j'entrave (le cheval)', *miel* 'agneau', *câine* 'chien', *(porcul) grohăie* '(le porc) grogne', *vier* 'verrat', *(găina) ciuguleşte* '(la poule) picore', *bărbie* 'caroncules (de la poule)'.

Nous formulons l'observation (importante dans ce contexte) que la réponse au questionnaire n'est pas toujours ce qu'on prévoit. Des fois, le répondant, ignorant la réalité visée par le questionnaire, s'en remet au paraverbal ou au nonverbal, s'il ne répond pas carrément par « je ne sais pas /je l'ignore». Dans d'autres cas, la réponse, quoiqu'elle s'actualise, peut être différente par rapport à ce qu'on anticipait, soit en tant que conséquence naturelle de la variation linguistique, soit comme résultat des limites que (surtout) les (jeunes) sujets ont dans la connaissance de certains termes spécialisés, désignant d'habitude des réalités vues plutôt comme archaïques. Les obstacles dans une comparabilité absolue des données en sont les conséquences les plus évidentes.

Enfin, on a sélectionné 26 questions du questionnaire, les réponses à celles-ci (28 pour chaque question : 7 points d'enquête x 4 informateurs) représentant un sous-ensemble de 728 séquences audio-vidéo, ce qui constitue pour nous le noyau initial à partir duquel nous envisageons de développer une ressource acoustique qui sera utilisé dans des recherches ultérieures.

Le traitement de la base de données

Dans la phase préliminaire, on a réalisé la conversion (à l'aide du logiciel *AVC free Converter*, pouvant être téléchargé de l'adresse http://www.any-video-converter.com/download-avc-free.php), à partir du format .flv en format .wav, d'un sous-ensemble sélectionné –

enregistrement contenant d'habitude un ou deux mots (le singulier et le pluriel d'un nom). Le traitement proprement-dit des données a porté sur deux aspects distincts, principalement sur la transcription phonétique du sous-ensemble sélectionné et, de l'autre côté, sur le traitement des enregistrements audio.

La transcription phonétique du sous-ensemble sélectionné

La transcription phonétique a été réalisée à l'aide du logiciel ALT_IIT, utilisé pour l'édition des troisième et quatrième volumes du NALR-Mold., Bucov.

Les traitements effectués en utilisant des fichiers audio

L'annotation du sous-ensemble audio sélectionné a été effectuée manuellement à l'aide de l'utilitaire Praat (cf. Boersma, Weenink 2010). À la suite de l'annotation manuelle on a obtenu pour chaque enregistrement un fichier texte ayant l'extension .TextGrid. Dans l'étape suivante, à l'aide d'un script, on génère pour chaque enregistrement sonore et pour le fichier .TextGrid correspondant, deux fichiers textes, le premier contenant les valeurs du formant F0 et le second contenant les valeurs des formants F1-F4. A l'aide des utilitaires HCopy et Hlist, ainsi que de HTK (http://htk.eng.cam.ac.uk/), pour chaque enregistrement on a généré deux fichiers textes (ayant l'extension .mfcc et .plp) contenant des valeurs des coefficients MFCC (Mel-fréquence Cepstral Coefficients), DMFCC (delta-MFCC), DDMFCC (delta-deltaMFCC), PLP (Perceptual Linear Prediction), DPLP (delta-PLP) et DDPLP (delta-delta-PLP).

La genèse des vecteurs de caractéristiques

Pour chaque enregistrement sonore, on peut générer plusieurs ensembles de caractéristiques à partir des valeurs des formants F0–F3, des coefficients MFCC et des coefficients DDMFCC DMFCC et PLP, DPLP et DDPLP (cf. Păvăloi, Muscă 2015). On a généré six ensembles de vecteurs de caractéristiques qui ont été utilisés par la suite pour la transcription phonétique automatique de [a], [e] et [i] du signal audio :

– SET1 – le premier ensemble de vecteurs de caractéristiques est obtenu en fonction des valeurs des 12 coefficients MFCC, à partir desquels on a génère les valeurs statistiques : la moyenne, la médiane et l'écart-type, puis le premier et le troisième quartile, parvenant en fin de compte à 60 caractéristiques (5 x 12);
– SET2 – le deuxième ensemble de vecteurs de caractéristiques est obtenu de manière analogue à la première série, sur la base des valeurs des coefficients MFCC, DMFCC respectivement DDMFCC;
– SET3 – le troisième ensemble de vecteurs de caractéristiques est obtenu de la même manière, à la différence que cette fois on utilise les 12 coefficients PL;
– SET4 – le quatrième ensemble de vecteurs de caractéristiques est obtenu de manière analogue à la deuxième série, sur la base des valeurs des coefficients PLP, DPLP et DDPLP, vecteurs ayant 180 caractéristiques;
– SET5 – est généré à partir des valeurs des formants F0–F3, en générant 45 caractéristiques statistiques pour chaque formant;
– SET6 – se compose de toutes les caractéristiques qui sont générées dans les cinq premières séries de traits.

Vu que les modèles génératifs GMM (Gaussian Mixture Model) (Gata, Toderean 2007) et HMM (Hidden Markov Model) (Gales, Young 2007) sont utilisés lorsqu'un grand nombre de vecteurs de caractéristiques est disponible (David 2002), ce qui n'est pas valable dans notre cas, on a utilisé deux classificateurs discriminants, k-NN (Dougherty 2013) et LS (Chang, Lin 2001). Dans les testes effectués à l'aide de k-NN on a utilisé différentes valeurs de k. Les résultats présentés ici sont valables pour des valeurs plus faibles de k, respectivement pour k = 1, k = 3 et k = 5. Les valeurs plus élevées de k ont généré globalement des résultats médiocres, aspect relativement normal étant donné qu'on a un nombre relativement restreint de vecteurs de caractéristiques. Cela peut changer pourtant avec l'augmentation du volume de l'ensemble de données. Dans les tests effectués à l'aide de l'algorithme k-NN on a utilisé cinq distances différentes : la

distance euclidienne, Manhattan, Canberra, Minkowsky, Cebychev.

Conception, traitement, résultats

La conception de l'architecture d'un logiciel système pour la transcription phonétique automatique

Après avoir déterminé toutes les exigences, on a conçu un modèle architectural qui comprend les étapes suivantes :

a) On crée la collection de données;
b) Pour chaque mot du dictionnaire contenant les transcriptions phonétiques (fichier ayant l'extension .DIC) :
b1) on parcourt l'annotation manuelle effectuée pour le mot envisagé;
b2) on lit les valeurs des formants, des coefficients MFCC et PLP;
b3) on génère, en fonction des options de l'utilisateur, les vecteurs caractéristiques – un vecteur pour chaque phonème;
c) On établit l'ensemble de données d'entraînement et de test à l'aide d'une technique de validation croisée (dans une première étape, on s'est proposé d'utiliser la technique LOO-CV – *Leave-one-out cross-validation*) ;
d) On utilise deux classificateurs discriminateurs, k-NN et SVM, et on accomplit la transcription phonétique automatique pour les voyelles /a/, /e/ et /i/;
e) On estime l'erreur de classification et on évalue le modèle à l'aide de la technique de validation croisée utilisée.

Résultats obtenus

En vue de l'interprétation des résultats obtenus par les classifications on a utilisé la matrice de confusion, étant donné que la qualité d'un classificateur, en termes d'identification correcte d'une classe, est mesurée à l'aide des informations fournies par la matrice de confusion. A partir de ces valeurs, on a calculé une série de mesures (Powers 2011) telles que Accuracy, Precision (*positive*

predictive value – PPV) et Recall (*true positive rate* – TPR ou Sensitivity), ce qui nous a permis d'interpréter les résultats obtenus par les classifications.

On présente ici une série de résultats préliminaires obtenus dans la transcription phonétique automatique pour les voyelles [a], [e] et [i].

En vue d'une transcription phonétique automatique pour les voyelles [a], [e] et [i] on a effectué des traitements en tenant compte de la base de données toute entière. On a effectué des tests en utilisant k-NN (k ayant les valeurs 1, 3 et 5) et SVM.

Pour la voyelle [a], le meilleur résultat, 73,10%, a été obtenu pour l'ensemble SET6 des vecteurs de caractéristiques, en utilisant k-NN (k = 1) et la distance Canberra. Pour la voyelle [e], le meilleur résultat, 84,62%, est obtenu pour l'ensemble SET1 des vecteurs de caractéristiques, en utilisant l'algorithme SVM. Le meilleur résultat obtenu à l'aide de l'algorithme k-NN, 79,25%, a été obtenu pour l'ensemble SET6 et k = 3, en utilisant les distances Manhattan et Canberra. Pour la voyelle [i], le meilleur résultat, 78,79%, est obtenu avec l'algorithme SVM. Le meilleur résultat auquel on est parvenu avec l'algorithme k-NN, 77,78%, a été obtenu pour l'ensemble SET6 pour k = 1 et les distances euclidienne et Canberra, ainsi que pour k = 5 et la distance Manhattan.

Commentaires, conclusions, perspectives

Commentaires

En ce qui concerne la reconnaissance du type du phénomène appliqué aux variables primaires des classes [a], [e] et [i], parce que le nombre de phénomènes phonétiques appliqués aux voyelles primaires est limité, les tests se sont avérés être irréalisables. Par exemple, pour la voyelle primaire [a] il n'y a que dix phénomènes:

– Trois phénomènes du groupe « Longueur » ;
– Trois phénomènes du groupe « Demi-longueur » ;
– Deux phénomènes du groupe « Assourdissement » ;
– Deux phénomènes du groupe « Fermeture ».

Leur nombre réduit rend impossible pour l'instant la réalisation de traitements dans le but d'identifier le type du phénomène phonétique appliqué aux voyelles primaires. Cependant, au fur et à mesure que la base de données se développera, ces tests deviendront réalisables.

Conclusions, perspectives

Puisqu'au début il était pratiquement impossible d'obtenir une collecte complète de données ayant un volume suffisant d'occurrences, on a choisi pour l'analyse seulement trois voyelles: [a], [e] et [i]. De toute évidence, dans les contextes analysés il y a aussi d'autres voyelles, mais les trois voyelles mentionnées sont prépondérantes. Pour parvenir à une approche plus complète du problème de la reconnaissance des voyelles, la base de données devra être mise à jour de manière adéquate, en ajoutant des transcriptions phonétiques pour des nouveaux enregistrements qui contiennent les voyelles visées.

De plus, à des fins expérimentales, on a sélectionné un informateur pour lequel on a effectué la transcription phonétique de toutes les réponses aux questions sélectionnées pour ALAB (jeune femme, point d'enquête Solca), afin de tester un système créé et entraîné pour un locuteur particulier. Le reste de la base des données a été réalisée à partir d'une série de questions dont les réponses ont été annotées pour tous les points d'enquête et pour chaque sujet.

Après les traitements effectués pour les enregistrements sélectionnés, on est arrivé à un certain nombre de constatations:

– le volume limité de la base de données existante rend relativement difficile une conclusion claire sur le pourcentage de la reconnaissance de la voyelle primaire pour les trois voyelles analysées – [a], [e] et [i] – en utilisant les algorithmes k-NN et SVM; cependant, les résultats pour ce qui est de la transcription phonétique automatique pour les voyelles [a] et [e] sont encourageants et à même de nous stimuler dans la poursuite de nos recherches;
– le fait qu'on n'a pas réalisé une validation de la transcription phonétique conduit de toute évidence à des pourcentages de

reconnaissance inférieurs à ce qui serait obtenu sur une collecte de données pour laquelle on aurait réussi une validation de la transcription phonétique. En d'autres termes, le biais de la transcription phonétique ajoute toute une série de difficultés supplémentaires au processus de la reconnaissance;

– le fait que certains pourcentages de reconnaissance sont plus élevés (même si seulement à quelques pourcents) que le pourcentage statistique de l'occurrence de la voyelle primaire étudiée suggère, cependant, la possibilité que la reconnaissance de la voyelle primaire se fasse automatiquement, avec une erreur acceptable pour la collecte des données validée;
– le problème de la reconnaissance du phénomène appliqué à une voyelle primaire ne peut être abordé que quand il y aura un minimum de données pour permettre des traitements de ce genre; on ne saurait pas prévoir un tel moment parce que, par exemple, même dans la situation d'un triplement de la base de données il est possible que le nombre de phénomènes appliqués ne soit que légèrement plus élevé que précédemment – ou qu'il augmente très peu;
– de toute évidence, dans l'analyse du texte on a la possibilité d'indiquer les semi-voyelles;
– la durée pourra être probablement établie en comparant la durée moyenne de la voyelle analysée avec la moyenne des durées pour la même voyelle primaire; les données existantes actuellement dans la base des données ne nous permettent pas encore d'accomplir cet objectif;
– le calcul et l'interprétation des mesures calculés à partir de la matrice de confusion mettent en évidence les voyelles primaires difficiles à reconnaître en raison de l'absence d'une quantité suffisante d'enregistrements dans la base de données; le problème est relativement simplifié, car on pourrait délimiter un ensemble de données dans lequel la probabilité de l'occurrence soit plus élevée et par conséquent, par la transcription phonétique de cette série d'enregistrements, on pourra augmenter de manière convenable la base de données existante.

Puisque l'analyse des résultats obtenus dans la reconnaissance de la voyelle primaire pour les timbres vocaliques [a], [e] et [i] permet des conclusions optimistes, on peut prévoir la poursuite de nos recherches, ce qu'on veut faire en plusieurs étapes:

– en premier lieu, l'augmentation de la base des données, qui doit devenir au moins trois ou quatre fois plus ample par rapport au volume actuel; il y a déjà un assez grand nombre (plus de 20% du volume total d'enregistrements sur le site ALAB) d'enregistrements annotés manuellement (dans la prochaine étape ça va grandir, pour atteindre au moins 33% de l'ensemble de données) et on se propose une augmentation significative des enregistrements transcrits phonétiquement;
– une éventuelle validation manuelle de la base de données (ou au moins d'un sous-ensemble qui remplisse le rôle des données d'entraînement); on considère même une validation automatique de la base de données (en réalisant un regroupement par clusters, par exemple, à l'aide de l'algorithme k-means, puis en validant un sous-ensemble de transcriptions phonétiques);
– on envisage une répétition du traitement de la nouvelle base de données (validée) pour les mêmes voyelles, [a], [e] et [i];
– dans la mesure où le volume total des données prétraitées (annotées manuellement et transcrites phonétiquement) le permettra, on va faire des traitements afin de reconnaître les voyelles primaires; à mentionner que pour la reconnaissance des voyelles primaires nous avons choisi une série de questions dont les réponses contiennent principalement les voyelles visées;
– on se propose l'analyse statistique du dictionnaire électronique correspondant au IIIème volume du NALR-Mold., Bucov., en déterminant les classes d'intérêt pour chaque voyelle (les voyelles primaires ayant un nombre significatif d'occurrences); de cette façon, on va éviter la situation qui s'est produite au cours du traitement de la base des données pour la voyelle [i], pour laquelle il y avait une classe [= phénomène phonétique] avec seulement deux occurrences, exemple à partir duquel on va déterminer un

seuil (un pourcentage minimum d'occurrences) qui produise des résultats supérieurs dans l'analyse;

– par la suite on va aborder éventuellement la question de la reconnaissance, en deux étapes: au début on doit reconnaître une classe plus générale (contenant plusieurs classes), tandis que dans la deuxième étape on va reconnaître la voyelle primaire.

En conclusion, on se propose pour la prochaine période l'augmentation du volume de la base des données et sa validation.

Bibliographie

APOPEI V., ROTARU F., BEJINARIU S., OLARIU F., *Electronic Linguistic Atlases*, (communication scientifique) « IKE'03, The 2003 Int'l Conference on Information and Knowledge Engineering » (Las Vegas, le 23 juin), 2003.

BEJINARIU S., APOPEI V., LUCA R., BOTOŞINEANU L., OLARIU F., *Instrumente pentru consultarea Atlasului Lingvistic şi editarea textelor dialectale*, in *Lucrările atelierului* ***Resurse lingvistice şi instrumente pentru prelucrarea limbii române*** (Iaşi, noiembrie 2006), Iaşi, Editura Universităţii « Alexandru Ioan Cuza », p. 107–112, 2006.

DOUGHERTY G., *Pattern Recognition and Classification – An Introduction*, New York, Springer Science and Business Media, 2013.

GALES M., YOUNG S., *The Application of Hidden Markov Models*, in *Speech Recognition, Foundations and Trends in Signal Processing*, vol. 1, no. 3, p. 195–304, 2007.

GATA M., TODEREAN G., *Results Obtained in Speaker Recognition Using Gaussian Mixed Models*, (communication scientifique) « The 4th Conference on Speech Technology and Human – Computer Dialogue (SpeD 2007) » (Iaşi, les 10–12 mai), 2007.

NALR-MOLD., BUCOV. : *Noul Atlas lingvistic român, pe regiuni. Moldova şi Bucovina*, vol. III, par Vasile Arvinte, Stelian Dumistrăcel, Ion A. Florea, Ion Nuţă, Adrian Turculeţ, Luminiţa Botoşineanu, Doina Hreapcă, Florin-Teodor Olariu, Iaşi, Editura Universităţii „Alexandru Ioan Cuza", 2007; vol. IV, par Vasile Arvinte, Stelian Dumistrăcel, Adrian Turculeţ, Luminiţa Botoşineanu, Doina Hreapcă, Florin-Teodor Olariu, Veronica Olariu, Iaşi, Editura Universităţii „Alexandru Ioan Cuza", 2014.

PĂVĂLOI I., MUSCĂ E., *Experimental Study in Development of Speech Corpus for Emotion Recognition with DataValidation*, (communication scientifique) « International Symposium on Signals, Circuits and Systems » (Iaşi, les 9–10 juillet), 2015.

POWERS D.M.W., *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation*, in « Journal of Machine Learning Technologies », no. 2 (1), p. 37–63, 2011.

Sources électroniques

BOERSMA P., WEENINK D., *Praat: doing phonetics by computer* (http://www.fon.hum.uva.nl/praat/), 2010.

CHANG C.-C., LIN C.-J., *Libsvm: a library for support vector machines*, version 2.3 (http://www.csie.ntu.edu.tw/~cjlin/ libsvm), 2001.

DAVID P.D., *Experiments with Speaker Recognition using GMM*, SpeechLab, Department of Electronics & Signal Processing, Technical University of Liberec, Hálkova, Czech Republic, Internal Report (https://www.scribd.com/ document/ 113088029/23), 2002.

http://www.any-video-converter.com/download-avc-free.php

http://htk.eng.cam.ac.uk/