Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Reconnaissance de parole beatboxée à l'aide d'un système HMM-GMM inspiré de la reconnaissance automatique de la parole

Solène Evain [1] Adrien Contesse [2] Antoine Pinchaud [3] Didier Schwab [1]
Benjamin Lecouteux [1] Nathalie Henrich Bernardoni [4]

(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
(2) http://www.vocalgrammatics.fr/
(3) ÉSAD Amiens, De-sign-e Lab, 80080 Amiens, France
(4) Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

solene.evain@univ-grenoble-alpes.fr, AdrienContesse@gmail.com,
APinchaud@gmail.com,Didier.Schwab@imag.fr, Benjamin.Lecouteux@imag.fr,
nathalie.henrich@gipsa-lab.fr

RÉSUMÉ

Le *human-beatbox* est un art vocal utilisant les organes de la parole pour produire des sons percussifs et imiter les instruments de musique. La classification des sons du beatbox représente actuellement un défi. Nous proposons un système de reconnaissance des sons de beatbox s'inspirant de la reconnaissance automatique de la parole. Nous nous appuyons sur la boîte à outils Kaldi, qui est très utilisée dans le cadre de la reconnaissance automatique de la parole (RAP). Notre corpus est composé de sons isolés produits par deux beatboxers et se compose de 80 sons différents. Nous nous sommes concentrés sur le décodage avec des modèles acoustiques monophones, à base de HMM-GMM. La transcription utilisée s'appuie sur un système d'écriture spécifique aux beatboxers, appelé Vocal Grammatics (VG). Ce système d'écriture s'appuie sur les concepts de la phonétique articulatoire.

ABSTRACT

BEATBOX SOUNDS RECOGNITION USING A SPEECH-DEDICATED HMM-GMM BASED SYSTEM [1]

Human beatboxing is a vocal art making use of speech organs to produce percussive sounds and imitate musical instruments. Beatbox sounds classification is a current challenge. We propose a beatbox sounds recognition system with an adaptation of the Kaldi toolbox, widely used for automatic speech recognition (ASR). Our corpus is composed of isolated sounds produced by two beatboxers and is composed of 80 different sounds. We focused on decoding with monophones acoustic models, trained with a HMM-GMM model. One type of transcription was used : a beatbox specific writing system named Vocal Grammatics (VG) which uses concepts of articulatory phonetics.

MOTS-CLÉS : human-beatbox, reconnaissance automatique de la parole, Kaldi, reconnaissance de sons isolés.

KEYWORDS: Human beatbox, automatic speech recognition, Kaldi, isolated sounds recognition.

1. Cet article a été publié en anglais dans le Workshop international MAVEBA http://maveba.dinfo.unifi.it/

Volume 1 : Journées d'Études sur la Parole, pages 208–216.
HAL : hal-02798538.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

1 Introduction

Le *human-beatbox* est apparu durant les années 80, dans le Bronx, un quartier de New York, et est associé à la culture hip-hop. Il consiste à produire des percussions vocales ainsi que des imitations d'instruments de musique, comme la trompette ou la guitare. La classification des sons de Beatbox peut être utilisée pour la recherche d'information musicale, comme une requête, pour la recherche de différents types de musique (Kapur *et al.*, 2004) ou pour des applications à commande-vocale avec un nombre de classes défini par l'utilisateur (Hipke *et al.*, 2014) afin de composer des morceaux de beatbox. Le beatbox est aussi utilisé dans le cadre de rééduction orthophonique : un système de reconnaissance peut permettre de travailler les exercices. Dans la littérature, des taux de classification corrects ont été obtenus sur un éventail limité de classes, c'est-à-dire cinq principaux sons de beatbox : *bass drum*, *open hi-hat*, *closed hi-hat*, *k-snare* et *p-snare* (Sinyor *et al.*, 2005). À notre connaissance, la reconnaissance automatique des sons de beatbox à l'aide d'un système de reconnaissance vocale n'a été explorée que par (Picart *et al.*, 2015). Leur base de données se compose de 5 sons percussifs de beatbox : *cymbal*, *hi-hat*, *kick*, *rimshot* , *snare* et 8 imitations d'instruments. Les performances étaient faibles pour les imitations d'instruments (taux d'erreur de reconnaissance de 41 %), mais plutôt correctes pour les classes se limitant aux sons percussifs (taux d'erreur de reconnaissance de 9 %).

En orientant nos efforts vers le développement d'un système de reconnaissance automatique des sons de beatbox efficace et fiable, nous visons à étendre le nombre de classes de sons et à permettre la reconnaissance de variantes subtiles dans la production de sons de beatbox. Nous considérons le *human-beatbox* comme un langage musical composé d'unités sonores que nous appellerons *boxèmes* en référence aux phonèmes de la parole. Par ailleurs, ce travail a été réalisé dans le but de créer un dispositif artistique interactif qui fournit des retours visuels lors de la production de sons de beatbox.

Le document est structuré comme suit : La section II présente la base de données. Le système de reconnaissance est présenté dans la section III. Différentes expériences sont décrites dans la section IV et leurs résultats sont donnés dans la section V. Les sections VI et VII présentent une discussion puis une conclusion, ainsi que les perspectives de nos travaux. Par ailleurs, ce travail a été présenté lors du workshop international MAVEBA (Evain *et al.*, 2019).

2 Corpus et matériel utilisé

Notre corpus de sons de beatbox appelé beatbox-VG2019 a été enregistré par deux beatboxeurs masculins : un beatboxeur professionnel (troisième auteur, nom de scène *Andro*) et un amateur (deuxième auteur). Il est composé de 80 boxèmes et peut être considéré comme un vaste corpus par rapport aux corpus précédemment présentés dans la littérature (et utilisés pour de la classification). Seuls les sons isolés sont considérés dans nos travaux, les séquences rythmiques étant écartées dans un premier temps.

Un système d'écriture pictographique basé sur l'articulation, développé par le deuxième auteur et s'appelant *Vocal Grammatics* (Contesse & Pinchaud, 2019) a été utilisé pour l'annotation. Dans cette écriture, les glyphes sont composés de deux informations : l'une sur les organes de la parole

Microphone	Distance de la bouche	Spécifications
Brauner VM1 (braun)	10 cm	condensateur + filtre pop
DPA 4006 (ambia)	50 cm	condensateur, micro d'ambiance
DPA 4060 (tie)	10 cm	condensateur
Shure SM58 (sm58p)	10 cm	dynamique
Shure SM58 (sm58l)	15 cm	dynamique
Shure beta 58 (beta)	1 cm	dynamique + encapsulé

TABLE 1: Récapitulatif des différents microphones

utilisés, l'autre sur la manière dont les sons sont produits (plosives, fricatives...). La Figure 1 illustre ce système d'écriture dans le cas d'un son plosif bilabial avec un glyphe morphologique représentant deux lèvres et un glyphe symbolique en forme de croix représentant la plosion.

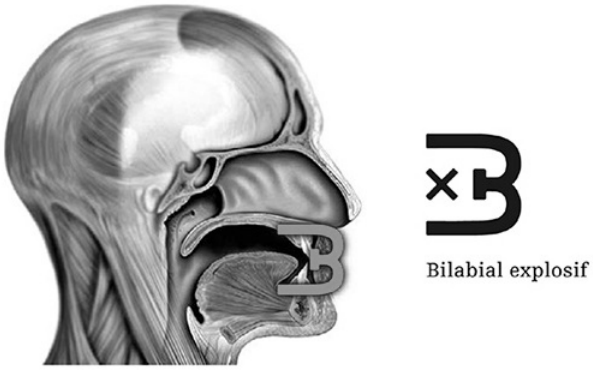


FIGURE 1: Représentation d'un son plosif bilabial avec le système d'écriture *Vocal Grammatics*.

Notre corpus de *boxèmes* a été enregistré avec six microphones. Cinq d'entre eux enregistraient simultanément et le dernier était encapsulé (une ou deux mains recouvrent la capsule du microphone). Les microphones différaient en termes de spécificités (par exemple, à condensateur ou dynamique) et de placement. Le tableau 1 donne les détails des microphones alors que le tableau 2 récapitule la composition du corpus.

L'apprentissage des modèles acoustiques a été réalisé avec la boîte à outils Kaldi (Povey *et al.*, 2011). En ce qui concerne les données de test, il s'agit de répétitions de différents sons de beatbox (pas toujours les même à la suite), dont la production est relativement lente (on peut percevoir une légère pause entre chaque son). Dans ce cadre de test, l'utilisation d'un système de reconnaissance automatique de la parole continue révèle son intérêt.

3 Reconnaissance du beatbox

Notre approche part de l'hypothèse que le *human-beatbox* est structuré comme un langage musical, utilisant les organes de la parole pour produire des unités sonores qui peuvent être distinguées les unes

Beatboxers	Adrien (amateur), Andro (professionnel)
Nom du corpus	beatbox-VG2019
Nombre de boxèmes (= taille du vocabulaire)	80
Nombre de boxèmes par beatboxeur	Adrien : 56/80 Andro : 80/80
Transcription	Vocal Grammatics
Microphone	5 simultanés + 1 encapsulé
Fréquence d'échantillonnage et précision	44100 Hz, 16 bits, mono
Durée totale d'enregistrement	∼206 min
Apprentissage	
Durée d'enregistrement	∼92 min
Nombre de répétitions des boxèmes	6 ou 2
Test	
Durée d'enregistrement	∼114 min
Nombre de répétitions des boxèmes	7 en moyenne

TABLE 2: Caractéristiques du corpus beatbox-VG2019

des autres et qui ont chacune une signification musicale spécifique pour le beatboxer. Dans ce contexte, un système de reconnaissance initialement dédié à la parole pourrait permettre de reconnaître les productions du beatbox. Généralement, dans le cadre de la reconnaissance automatique de la parole, les mots sont décomposés en unités (phonèmes, syllabes etc.) qui permettent de définir un lexique associant chaque mot à sa représentation sous forme d'unités atomiques. Les modèles acoustiques sont alors entraînés pour reconnaître ces unités.

Dans ce travail préliminaire, nous avons considéré chaque son de beatbox comme étant atomique : nous partons sur une approche de reconnaissance de mots isolés. La co-articulation ou la frontière extra-boxèmes ont été écartées, tout en conservant les contraintes de traitement du bruit, de variabilité intra et inter-locuteur. À terme nous souhaitons travailler sur des sous-unités au niveau des sons, mais nous manquons encore de données pour généraliser ces sous-unités.

Les paramètres utilisés sont de type MFCC. Ces paramètres se basent sur le système auditif périphérique humain (Tiwari, 2010) et sont largement utilisées dans les systèmes de reconnaissance automatique de la parole. Chaque son de beatbox a été associé à un modèle de Markov caché (HMM). Nous nous sommes limités à une approche HMM-GMM car les quantités de données d'apprentissage sont très faible (quelques dizaines de minute) et la parole beatboxée est tellement spécifique qu'il nous semblait difficile d'exploiter des méthodes neuronales (par apprentissage direct ou même par transfert d'apprentissage).

Notre objectif est avant tout d'appliquer les principes d'un système de reconnaissance de la parole continue avec la constitution d'unités acoustiques (nos boxèmes), d'un lexique (nos 80 sons pour l'instant) et d'un modèle de langage (qui n'est pas traité ici, mais qui représenterait la rythmique des séquences beatboxées).

Dans nos expériences, l'apprentissage a été réalisé avec une sélection du nombre de Gaussiennes automatique en fonction de la quantité de données (cependant, nous avons essayé différentes quantités de Gaussiennes, sans observer le moindre impact). Par ailleurs, au niveau du modèle de langage,

la probabilité d'émission d'un boxème est identique pour chacun d'eux étant donné que dans ces expériences préliminaires nous ne prenons pas en compte les séquences. Le système est donc capable de reconnaître de manière continue les boxèmes produits sans avoir de connaissance *a priori* sur la rythmique.

4 Méthode

Plusieurs systèmes ont été conçus dans le but de tester divers paramètres. L'influence dans l'apprentissage de chaque microphone avec différents placements et sensibilités a été étudiée afin de savoir si tous les enregistrements pouvaient être utilisés ensemble pour former un système plus robuste. Pour chaque microphone, nous avons découpé les enregistrements en une base d'apprentissage (6 répétitions de boxèmes) et une base de test (1 à 12 répétitions de boxèmes). Les résultats nous permettent de classer les microphones du plus efficace au moins efficace.

Pour cette première étude, nous avons souhaité nous concentrer sur les boxèmes produits de façon non-encapsulée. Le fait d'encapsuler le microphone (le recouvrir de la main) modifie le résultat sonore pour un boxème donné. Comme cinq des six microphones étudiés ont été utilisés de façon non-encapsulée par les beatboxeurs, le nombre d'enregistrements disponibles est plus conséquent que celui des boxèmes produits de façon encapsulés.

L'impact de différents paramètres sur la reconnaissance a été testé : le nombre d'états des HMM, les paramètres MFCC, la probabilité d'apparition d'un silence, ainsi que l'ajout ou non d'un phonème de pause dans le lexique . Certains choix ont été basés sur l'article de (Picart *et al.*, 2015). Nous présentons ici les résultats pour quatre configurations du système de reconnaissance :

- configuration A : les quatre paramètres cités ci-dessus sont par défaut, à savoir 3 états HMM, 13 paramètres MFCC, la probabilité d'apparition d'un silence à 0.5 et l'absence de phonème pause dans le lexique ;
- configuration B : une pause a été ajoutée dans le lexique et la probabilité d'apparition d'un silence est fixée à 0.8 ;
- configuration C : même base que la configuration B. Le nombre de coefficients MFCC passe à 22 ;
- configuration D : même base que la configuration B. Le nombre d'états HMM passe à 5.

Le lexique d'un système de reconnaissance de la parole est de la forme 'mot : transcription phonétique'. Ici, le mot est un boxème. La pause indiquée dans les systèmes ci-dessus a été ajoutée dans le lexique de la façon suivante : 'boxème : pause transcription_phonétique pause'. Cette pause n'est pas présente dans la transcription manuelle des corpus de test et n'est pas présente dans l'hypothèse de décodage. Les systèmes B, C et D sont donc comparables au système A puisque la valeur de dénominateur du BER est la même.

La mesure d'évaluation 'BER' -*Boxeme Error Rate*- est utilisée pour évaluer le système. Elle est directement inspirée du taux d'erreur sur les mots (WER) puisqu'il est calculé en additionnant le nombre de substitutions, d'insertions et de suppressions divisé par le nombre de boxèmes dans la référence. Plus la reconnaissance est bonne, plus la valeur du BER est faible. Le CBR (taux de boxèmes correct) est également utilisé en table 3 afin d'avoir une deuxième mesure de l'efficacité du système. Il indique le pourcentage de boxèmes correctement reconnus.

5 Résultats

Les figures 2 et 3 donnent le BER pour différents décodages. La ligne "but" sur l'axe horizontal représente notre objectif : obtenir un BER de 10 % ou moins, *a priori* fixé pour garantir une utilisation intéressante de notre système par le public.

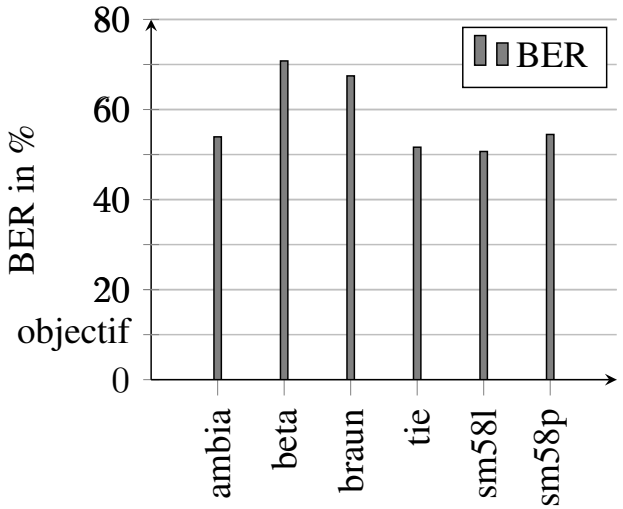


FIGURE 2: BER obtenu avec des modèles acoustiques monophones pour les six microphones

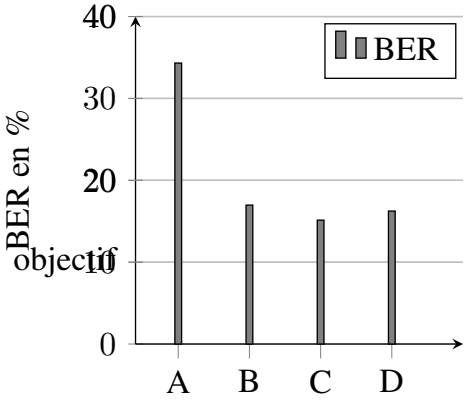


FIGURE 3: Évolution du BER pour les configurations A, B, C et D du système
A : par défaut / **B** : prob. silence=0.8 + pause / **C** : B + MFCC=22 / **D** : B + états HMM=5

En figure 2 nous pouvons observer l'efficacité de chacun des six microphones. Nous avons constaté que les microphones à condensateur DPA et les microphones dynamiques Shure SM58, placés à proximité ou loin de la bouche du beatboxer, offrent des performances similaires. Des taux de reconnaissance plus faibles sont constatés pour les enregistrements avec le microphone dynamique Shure beta 58 encapsulé et le microphone à condensateur Brauner VM1.

Nous avons ensuite voulu tester l'incidence de certains paramètres sur la reconnaissance. Nous avons sélectionné un microphone pour ces tests : le Shure SM58 - 'sm58p' - utilisé proche de la bouche. Notre choix s'est porté sur celui-ci particulier car c'est un des microphones le plus utilisé par les beatboxeurs à l'échelle mondiale et que l'utilisation du microphone dans cette discipline est le plus souvent effectuée proche de la bouche. Un premier test a été de faire varier la probabilité de silence de 0,5 (par défaut) à 0,9 avec un pas fixé à 0,1. Notre meilleur modèle a été obtenu avec une probabilité de silence de 0,9, ce qui a donné un BER de 26,94 %. En spécifiant en plus un phonème de "pause" avant et après chaque boxème dans le lexique, nos meilleurs résultats sont de 16,97% de BER. Ceux-ci sont obtenus avec une probabilité de silence de 0,8. Cela s'explique par la configuration de nos expériences : nous avons demandé aux beatboxers de répéter un même son ; des pauses ont donc artificiellement été ajoutées en début ou fin de chaque son.

La figure 3 montre l'évolution des résultats avec différents paramètres : probabilité de silence plus élevée, ajout d'une pause dans le lexique, 22 paramètres MFCC au lieu de 13 par défaut et 5 états par HMM au lieu de 3 par défaut. Pour rappel, la configuration A est représentative d'un système avec les quatre paramètres laissés 'par défaut'. L'ensemble d'apprentissage a été réalisé avec des enregistrements de microphones non encapsulés.

Notre meilleur modèle est obtenu avec la configuration C et donne un BER de 15,13 %. Les configurations B et D sont très proches avec des BER de 16,97% et 16,24% respectivement (voir tableau 3 pour les détails concernant les substitutions, les insertions, les suppressions et les taux de boxèmes corrects).

Dans la figure 3 et le tableau 3, nous observons que chaque changement de paramétrage est bénéfique pour les substitutions, les insertions, les suppressions et les taux de boxèmes corrects. Le bénéfice le plus évident est pour le taux d'insertion qui passe à zéro. Le taux de boxème correct atteint 85% avec la configuration C.

	A	**B**	**C**	**D**
Substitutions	19.19%	12.73%	10.70%	12.36%
Insertions	9.41%	0.18%	0.18%	0%
Deletion	5.72%	4.06%	4.24%	3.87%
CBR	75.09%	83.21%	85.06%	83.76%

TABLE 3: Insertions, substitutions, suppressions et taux de boxème correct (CBR) pour les configurations A B C D
A : par défaut / **B** : probabilité de silence 0.8 + pause / **C** : B + 22 MFCC / **D** : B + HMM à 5 états

6 Discussion

Comme nous l'avons vu précédemment, l'impact des différents microphones est assez faible, à l'exception des microphones Shure beta 58 et Brauner VM1 qui sont moins performants. Nous supposons que c'est à cause de la façon dont nous les avons utilisés (proximité par rapport au locuteur). En effet, le microphone Shure beta 58 est encapsulé et cette utilisation affecte les performances du microphone. Quant au microphone à condensateur Brauner VM1, nous pouvons observer qu'il fonctionne moins bien que l'autre microphone à condensateur de notre test (DPA 4060) et supposons qu'il a été placé trop près de la bouche du beatboxer. Enfin, ni le nombre de paramètres MFCC ni le nombre d'états dans le HMM n'apportent une nette amélioration. Nous supposions qu'augmenter le nombre d'états HMM était intéressant pour les sons complexes qui sont composés de deux ou plusieurs boxèmes. Ces aspects seront analysés dans des études ultérieures.

7 Conclusion et perspectives

Notre approche démontre qu'utiliser un système de reconnaissance vocale pour reconnaître les sons de beatbox isolés est pertinent. Cela ouvre des perspectives pour la reconnaissance vocale de phrases beatboxées.

Jusqu'à présent, notre meilleur modèle a été obtenu avec une augmentation de la probabilité de silence (0,8 au lieu de 0,5), l'insertion d'un phonème de silence "pause" étant ajouté dans les contextes droits et gauche du vocabulaire et 22 paramètres pour les MFCC. Le meilleur BER obtenu est alors de 15,13%.

Nous avons pu observer que le type de microphone utilisé pour l'enregistrement ne semble pas avoir

d'influence sur le système. Il dépend plutôt de leur utilisation (encapsulé ou non). Mettre de côté le microphone encapsulé pour l'apprentissage donne de meilleurs résultats.

Quant aux différents types de production, lorsqu'ils sont mélangés, ils semblent dégrader fortement les performances. Pour l'instant, en ce qui concerne les substitutions, nous ne pouvons rien conclure car le système semble mélanger des sons qui sont assez semblables à l'oreille ou qui ont une articulation assez similaire, et des sons qui sont très différents. Nous supposons que la division du corpus en fonction de la longueur du son et l'adaptation du nombre d'états HMM pourraient améliorer le système.

Diviser chaque son en plus petits morceaux, comme on le fait pour les langues comportant des phonèmes ou des syllabes, est une perspective. En effet, à mesure que le vocabulaire du corpus augmentera, nous serons confrontés à un manque d'exemples pour l'apprentissage. Le fait de disposer d'un modèle basé sur des boxèmes réduirait le nombre de modèles nécessaires au système et permettrait le traitement de la coarticulation. De plus, il reste à explorer les séquences rythmiques (que nous pourions apparenter à un modèle de langage) et la reconnaissance des sons encapsulés. Enfin, il serait intéressant de voir si la reconnaissance des voix des femmes ou des enfants pose des problèmes dans le cadre des sons de beatbox.

Des perspectives plus techniques visent à résoudre le fait que les données annotées de beatbox sont pour l'instant très précieuses et rares. En effet, dans les expériences décrites nos ensembles d'apprentissage ne représentent tout au plus que quelques dizaines de minutes. Pour cette raison nous sommes restés concentrés sur des modèles de type HMM-GMM qui sont moins gourmands en données que des modèles à base de réseaux de neurones profonds : nous envisageons d'exploiter des techniques d'augmentation de données, de synthèse de données et l'utilisation d'outils non supervisés tels que wav2vec.

Références

CONTESSE A. & PINCHAUD A. (2019). *vocal grammatics*. Web page, www.vocalgrammatics.fr, Last consulted : 2019-08-29.

EVAIN S., CONTESSE A., PINCHAUD A., SCHWAB D., LECOUTEUX B. & HENRICH BERNARDONI N. (2019). Beatbox sounds recognition using a speech-dedicated hmm-gmm based system.

HIPKE K., TOOMIM M., FIEBRINK R. & FOGARTY J. (2014). BeatBox : End-user Interactive Definition and Training of Recognizers for Percussive Vocalizations. p. 121–124, Como, Italy : ACM.

KAPUR A., TZANETAKIS G. & BENNING M. (2004). Query-by-Beat-Boxing : Music Retrieval For The DJ. Barcelona, Spain.

PICART B., BROGNAUX S. & DUPONT S. (2015). Analysis and automatic recognition of Human BeatBox sounds : A comparative study. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 4255–4259, Brisbane, QLD, Australia. DOI : 10.1109/ICASSP.2015.7178773.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P., SILOVSKY J., STEMMER G. & VESELY K. (2011). The Kaldi Speech Recognition Toolkit. p.4̃, Hilton Waikoloa, Big Island, Hawaii, US.

SINYOR E., MCKAY C., FIEBRINK R., MCENNIS D. & FUJINAGA I. (2005). Beatbox classification using ACE. p.4̃, London, UK.

TIWARI V. (2010). MFCC and its applications in speaker recognition. *International Journal on Emerging Technologies*, p. 19–22.