Transformation de la prosodie par adaptation MLLR de GMM

Damien Lolive, Nelly Barbot, Olivier Boeffard

IRISA / Université de Rennes 1 - ENSSAT
6 rue de Kerampont, B.P. 80518, F-22305 Lannion Cedex
France
{damien.lolive,nelly.barbot,olivier.boeffard}@irisa.fr

ABSTRACT

In a voice transformation context, prosody transformation using parallel corpora is quite unrealistic as such corpora are difficult and also expensive to build. Based on this observation, we propose an approach for transforming prosody using non-parallel corpora thanks to the MLLR adaptation strategy. This methodology is applied to the joint transformation of duration and F_0 at the syllable level. The source data are modelled by a GMM which is adapted to the target by applying a linear transformation to the mean vectors of the gaussian mixture. This methodology is applied to the conversion of duration and F_0 between two french speakers and is evaluated by cross validation between the models and the test datasets.
Keywords: prosody transformation, non-parallel corpora, GMM, MLLR

1. Introduction

Un système de transformation de la voix a pour objectif de modifier les phrases d'un locuteur source pour qu'elles soient perçues comme si elles avaient été prononcées par un locuteur cible. Au cours de ces dernières années, certains domaines technologiques comme l'identification biométrique ou encore la synthèse de la parole à partir du texte font usage d'une méthodologie de transformation de la voix. En ce qui concerne l'identification biométrique, les transformations de la prosodie et de la voix, en général, peuvent être utilisées pour tester des systèmes de vérification ou d'identification du locuteur. En ce qui concerne la synthèse de parole, la transformation de voix peut avoir un impact important dans la mesure où, un corpus d'unités acoustiques décrivant une voix de synthèse, associé à un ensemble de fonctions de transformation, pourrait se substituer à l'approche classique qui nécessite un corpus différent pour chaque voix.

Plus précisement, une telle fonction doit réaliser à la fois une transformation des caractéristiques acoustiques segmentales et supra-segmentales (prosodie). Dans cet article, nous nous plaçons dans le cadre de la transformation de la prosodie en considérant les facteurs de la durée et de la fréquence fondamentale, F_0. Un tel système de transformation peut se décomposer en trois phases : stylisation, classification puis transformation. Dans la littérature, un nombre important de travaux récents traitent de la transformation de la prosodie et plus particulièrement du F_0 [9, 7]. Une approche classique consiste à modifier les contours de F_0 par une transformation linéaire ou polynômiale, qui repose sur des paramètres globaux du F_0 de la source et de la cible [3, 2]. D'autres approches décomposent le problème complexe de transformation en sous-problèmes par un partitionnement de l'espace, comme par exemple avec un *codeboook* [3, 8].

Dans les systèmes de conversion de voix classiques, il est nécessaire d'avoir pour chaque phrase, un exemplaire de la source et un autre de la cible. En conséquence, deux corpus *parallèles* doivent être utilisés, ce qui constitue une hypothèse restrictive, pas toujours applicable selon l'application désirée. Relâcher cette contrainte permettrait de concevoir des applications de manière plus souple. Une solution possible à ce problème, sous l'hypothèse de modèles paramétriques, est l'adaptation de ces modèles via une technique d'adaptation au locuteur comme la MLLR, Maximum Likelihood Linear Regression, [6].

Nous proposons de mettre en œuvre une technique de transformation de la prosodie dans le cadre de corpus non parallèles. Nous traitons l'information prosodique au niveau de la syllabe. L'idée sous-jacente est qu'une phrase mélodique peut être décomposée en éléments de plus petite taille pouvant être assemblés pour former une phrase mélodique complète [11]. En suivant cette hypothèse, la transformation de la prosodie est réalisée par séquences syllabiques. Au niveau d'une syllabe, la durée et le F_0 sont représentés sous la forme d'un vecteur de taille fixe. Nous avons choisi de modéliser l'espace mélodique d'un locuteur par un GMM sur ces vecteurs de taille fixe. L'adaptation des paramètres du GMM source avec les données de la cible est ensuite obtenue en appliquant une méthodologie MLLR. Une fonction de transformation utilisant les paramètres du GMM adapté est également proposée. Le vecteur transformé est construit comme une pondération des centroïdes du GMM source. Les coefficients de pondération correspondent à la probabilité a posteriori de considérer une composante du GMM après observation du vecteur source. Cette approche a déjà été proposée pour la transformation du timbre et a montré une meilleure efficacité de transformation que l'approche mapping code-book [13].

La représentation de la durée et du F_0 est décrite dans la section 2. La section 3 présente la modélisation GMM utilisée ainsi que l'adaptation des paramètres aux données du locuteur cible. La fonction de transformation est également détaillée dans cette partie. La méthodologie expérimentale est ensuite présentée

en section 4. Les résultats sont donnés et discutés en section 5.

2. Prétraitement des données

Les contours du F_0 au niveau de la phrase sont prétraités comme proposé par [14]. Une interpolation est d'abord réalisée pour éliminer les parties non voisées de la courbe de F_0. Cette interpolation résulte de l'hypothèse selon laquelle il existe un geste mélodique continu. Les valeurs de la fréquence fondamentale seraient alors masquées sur des segments non voisés. De plus, les contours de F_0 obtenus sont lissés grâce à une spline cubique afin de supprimer les variations microprosodiques. La prosodie d'une syllabe (F_0 et durée) est représentée par un vecteur de dimension 6 :

$$\mathbf{x} = (d_{\text{onset}}, d_{\text{nucleus}}, d_{\text{coda}}, F_0^{10\%}, F_0^{50\%}, F_0^{90\%})$$

Ce vecteur permet de traiter de façon conjointe la transformation de la durée et du F_0. La représentation de la durée repose sur la structure d'un contour de F_0 au niveau syllabique : onset, nucleus et coda. La durée de chaque zone est calculée en multiple de 10 ms. Concernant le F_0, on résume le contour à trois points cibles situés à 10%, 50% et 90% du support. L'utilisation d'une telle représentation revient à normaliser la durée d'un contour par rapport à un support temporel unique pour tous les contours comme cela est effectué dans [12]. Cela permet d'éliminer les variations de longueur des contours de F_0 tout en permettant une comparaison de la forme des contours.

3. Transformation de la prosodie

Le but de notre approche est de transformer la prosodie entre un locuteur source et un locuteur cible sans qu'il soit nécessaire d'utiliser des corpus parallèles. Des GMM sont utilisés pour modéliser les vecteurs sources et cibles et la fonction de conversion repose sur les paramètres du GMM adapté source/cible. La première phase consiste à apprendre un GMM sur les données du locuteur source puis à adapter ses paramètres avec les données du locuteur cible. La seconde phase du système concerne l'utilisation des modèles pour effectuer la transformation de la durée et du F_0.

3.1. Modélisation GMM

On considère pour un locuteur l'ensemble $\mathbf{X}$ des vecteurs $\mathbf{x}$ représentant la prosodie de chaque syllabe. Un modèle GMM $\mathcal{M}_{\mathbf{X}}$ à M composantes est choisi pour modéliser l'ensemble $\mathbf{X}$, sa distribution de probabilité est donnée par :

$$P(\mathbf{x}|\mathbf{\Theta}) = \sum_{m=1}^{M} \alpha_m P(\mathbf{x}|\theta_m) \quad (1)$$

où le vecteur de paramètres du modèle est $\mathbf{\Theta} = (\alpha_1, \ldots, \alpha_m, \theta_1, \ldots, \theta_M)$. α_m est le coefficient de mélange associé à la m-ème gaussienne de paramètre $\theta_m = (\mu_m, \Sigma_m)$ et de distribution $P(\mathbf{x}|\theta_m)$.

L'algorithme EM a été mis en œuvre pour apprendre les GMM en maximisant la log-vraisemblance des données et du modèle, [1].

3.2. Adaptation d'un GMM

La méthode d'adaptation MLLR (Maximum Likelihood Linear Regression) est proposée dans [6]. Considérons un GMM $\mathcal{M}_{\mathbf{X}}$ avec comme vecteur de paramètres $\mathbf{\Theta} = (\alpha, \mu, \mathbf{\Sigma})$ appris sur l'ensemble de données $\mathbf{X}$. Le but est d'adapter les paramètres du GMM $\mathcal{M}_{\mathbf{X}}$ à l'ensemble de données $\mathbf{Y}$ en calculant pour chaque composante de $\mathcal{M}_{\mathbf{X}}$ une transformation linéaire de ses paramètres μ_m et $\mathbf{\Sigma}_m$ de manière à maximiser la vraisemblance du GMM adapté sur l'ensemble $\mathbf{Y}$:

$$\begin{aligned} \hat{\mu}_m &= \widehat{\mathbf{W}}_m \xi_m \\ \widehat{\mathbf{\Sigma}}_m &= \mathbf{B}_m^T \widehat{\mathbf{H}}_m \mathbf{B}_m \end{aligned}$$

où $\widehat{\mathbf{H}}_m$ est la matrice de transformation de la variance et $\mathbf{B}_m$ est l'inverse du facteur de Choleski de $\mathbf{\Sigma}_m^{-1}$. $\widehat{\mathbf{W}}_m$ est la matrice d'adaptation de la moyenne de taille $n \times (n+1)$ (n est la taille d'un vecteur de données) et $\xi_m = [1 \mu_1 \ldots \mu_n]$ est le vecteur des moyennes étendu.

L'approche MLLR consiste à trouver un ensemble de matrices de transformation qui, lorsqu'elles sont appliquées aux moyennes et variances des gaussiennes, permettent de maximiser la vraisemblance des données d'adaptation. L'estimation de $\widehat{\mathbf{W}}_m$ et de $\widehat{\mathbf{H}}_m$ est réalisée en appliquant l'algorithme EM avec les données d'adaptation.

3.3. Transformation de la durée et du F_0

La prosodie de la voix source doit être modifiée afin de ressembler à la prosodie de la voix cible. Par analogie avec les travaux effectués dans le domaine du segmental [13], on définit la fonction de transformation de la manière suivante :

$$\mathbf{x}' = \mathcal{F}(\mathbf{x}, \mathcal{M}_{\mathbf{X}}) = \sum_m P(m|\mathbf{x}) \mu_m \quad (2)$$

où $P(m|\mathbf{x})$ est la probabilité que $\mathbf{x}$ appartienne à la classe m et μ_m est la moyenne de la gaussienne m du GMM $\mathcal{M}_{\mathbf{X}}$.

En particulier, pour transformer les vecteurs de la source vers l'espace des vecteurs de la cible, on utilise le GMM adapté $\widehat{\mathcal{M}_{\mathbf{X}}}$. La fonction de transformation s'écrit en utilisant la matrice de transformation calculée lors de l'adaptation MLLR :

$$\mathcal{F}(x, \widehat{\mathcal{M}_{\mathbf{X}}}) = \sum_m P(m|\mathbf{x}) \hat{\mu}_m = \sum_m P(m|\mathbf{x}) (\widehat{\mathbf{W}}_m \xi_m) \quad (3)$$

où ξ_m est le vecteur moyen étendu de la classe m du GMM source $\mathcal{M}_{\mathbf{X}}$, $\widehat{\mathbf{W}}_m$ est la matrice d'adaptation pour cette gaussienne.

4. Protocole expérimental

4.1. Données

Pour réaliser les expérimentations, deux corpus de données issus de BREF120 [10] sont utilisés. BREF120 est un corpus de français multi-locuteur. La sélection d'un couple de voix parmi toutes les combinaisons possibles a été réalisée grâce à un test subjectif de dissemblance de la prosodie. Ce test a été

Tab. 1: Log-vraisemblances pour $\mathcal{M}_{\mathbf{X}}$ et $\mathcal{M}_{\mathbf{Y}}$ ainsi que les résultats de l'adaptation de $\mathcal{M}_{\mathbf{X}}$ en utilisant $\mathbf{Y}$ avec un intervalle de confiance à 95%.

	Appr.	Valid.
$\mathcal{M}_{\mathbf{X}}$	-21.11 ± 0.10	-21.25 ± 0.19
$\mathcal{M}_{\mathbf{Y}}$	-20.74 ± 0.10	-20.86 ± 0.20
$\widehat{\mathcal{M}_{\mathbf{X}}}$	-21.08 ± 0.10	-21.22 ± 0.18

Tab. 2: Log-vraisemblances pour les GMM source $\mathcal{M}_{\mathbf{X}}$, cible $\mathcal{M}_{\mathbf{Y}}$ et adapté $\widehat{\mathcal{M}_{\mathbf{X}}}$ sur les données de validation source $\mathbf{X}$ et cible $\mathbf{Y}$.

	X	Y
$\mathcal{M}_{\mathbf{X}}$	-21.25 ± 0.19	-21.65 ± 0.19
$\mathcal{M}_{\mathbf{Y}}$	-21.49 ± 0.22	-20.86 ± 0.20
$\widehat{\mathcal{M}_{\mathbf{X}}}$	-21.68 ± 0.20	-21.22 ± 0.18

effectué sur les 40 couples de voix du même genre possédant le plus de phrases communes. 9 auditeurs ont répondu à la question "Le style d'élocution (intonation, débit, etc.) de ces deux voix vous semble-t-il : dissemblables, ..., identiques" (sur une échelle de 0 à 4). Les réponses ont permis de retenir les voix de deux femmes avec un score de $0,43$: *JMF* et *JNF*, resp. la source et la cible.

Pour les deux corpus, la segmentation en phones est réalisée de manière automatique [4]. La fréquence fondamentale moyenne, F_0 a été obtenue grâce à la méthode YIN [5]. Chaque séquence phonétique est segmentée en syllabes. Les contours de F_0 sont interpolés puis lissés avant de constituer les vecteurs de données comme cela est présenté en section 2. Pour chaque voix, 5000 syllabes sont utilisées pour l'apprentissage des GMM et 1500 syllabes pour la validation.

4.2. Expériences

Les GMM utilisés pour ces expériences possèdent 32 gaussiennes avec des matrices de covariance diagonales. Le processus d'adaptation, ainsi que la transformation du F_0 et de la durée sont difficiles à évaluer. En effet, en respectant un cadre de travail strictement non parallèle, des tests subjectifs sont difficilement praticables. Nous proposons ici une méthodologie d'évaluation par validation croisée reposant sur les log-vraisemblances entre les modèles et les données. Nous considérons les GMM source $\mathcal{M}_{\mathbf{X}}$ et cible $\mathcal{M}_{\mathbf{Y}}$ respectivement appris sur les données source et cible. On définit également le GMM adapté $\widehat{\mathcal{M}_{\mathbf{X}}}$ appris d'abord sur les données source et ensuite adapté par rapport aux données cibles. La fonction de transformation des données source en utilisant le modèle adapté est décrite dans (3). De plus, pour évaluer la qualité des données transformées par rapport aux données cibles, il est nécessaire d'évaluer tout d'abord l'impact de la fonction de transformation (2) sur les données initiales.

5. Résultats et discussion

Dans le tableau 1, nous pouvons observer les valeurs de log-vraisemblance pour l'apprentissage des GMM source et cible ainsi que pour l'adaptation du GMM source aux données cibles. Dans tous les cas, les valeurs de log-vraisemblance pour l'apprentissage sont proches des valeurs pour la validation. Ces résultats montrent que les GMM sont bien adaptés aux données et aucun effet d'overfitting n'apparaît.

Le tableau 2 montre une comparaison des valeurs de vraisemblance pour les GMM source $\mathcal{M}_{\mathbf{X}}$, cible $\mathcal{M}_{\mathbf{Y}}$ et adapté $\widehat{\mathcal{M}_{\mathbf{X}}}$ par rapport respectivement aux données source $\mathbf{X}$ et cible $\mathbf{Y}$. En analysant ce tableau ligne par ligne, on peut noter que les GMM $\mathcal{M}_{\mathbf{X}}$ et $\mathcal{M}_{\mathbf{Y}}$ ont de meilleurs résultats respectivement sur les voix source et cible. De plus, le GMM $\widehat{\mathcal{M}_{\mathbf{X}}}$ est mieux adapté aux données de la cible, $\mathbf{Y}$, qu'à celles de la source, $\mathbf{X}$. Ces résultats montrent que le processus d'adaptation MLLR a déplacé les distributions des gaussiennes de $\mathcal{M}_{\mathbf{X}}$ vers celles de $\mathcal{M}_{\mathbf{Y}}$. Ce mouvement peut être observé sur la figure 1. On peut observer que la distribution adaptée 1(b) est plus proche de la distribution des données cibles 1(c). En effet, l'adaptation de la variance a permis d'élargir certaines gaussiennes tandis que l'adaptation de la moyenne a permis de déplacer la distribution source 1(a) vers la droite.

Afin d'évaluer le comportement de la fonction de transformation, considérons $\mathbf{X}'$, $\mathbf{Y}'$, $\mathbf{Z}'$ trois nouveaux ensembles de données obtenus par transformation respectivement des données :
– source par le GMM source $\mathcal{M}_{\mathbf{X}}$, en appliquant (2),
– cible par le GMM cible $\mathcal{M}_{\mathbf{Y}}$, en appliquant (2),
– source par le GMM adapté $\widehat{\mathcal{M}_{\mathbf{X}}}$, en appliquant (3).
Les résultats obtenus dans le tableau 3 sont meilleurs qu'avec les ensembles de données originaux $\mathcal{M}_{\mathbf{X}}$ et $\mathcal{M}_{\mathbf{Y}}$. L'explication de ce phénomène est que la fonction de transformation a tendance à projeter les données vers les moyennes des gaussiennes. En effet, l'équation (2) indique que la valeur transformée est égale à la somme des moyennes des gaussiennes pondérées par la probabilité d'appartenance à une classe. Plus précisément, les données transformées sont situées dans l'enveloppe convexe formées par les moyennes des gaussiennes du GMM. De ce fait, la variance des données transformées est uniquement liée à la probabilité d'appartenance à une classe et elles possèdent donc une variance plus faible que les données originales. Des expériences complémentaires, non présentées ici par manque de place, ont permis de confirmer ce comportement de la fonction de transformation. Pour remédier à ce manque de variabilité, il serait utile d'introduire la variance des gaussiennes dans la fonction de tranformation.

Les résultats obtenus pour le GMM $\mathcal{M}_{\mathbf{Y}}$ montrent que les données adaptées ressemblent plus aux données cibles transformées $\mathbf{Y}'$ qu'aux données sources transformées $\mathbf{X}'$. Pour le GMM $\widehat{\mathcal{M}_{\mathbf{X}}}$, les résultats sont plus mitigés. En effet, il semble que les données $\mathbf{X}'$ soient plus vraisemblables que les données $\mathbf{Y}'$. Cela a tendance à montrer que les données source et cible, sont difficilement séparables. Il serait alors intéressant d'enrichir le modèle de prosodie, par exemple en introduisant la dérivée du F_0 au niveau des trois points sélectionnés. Ces informations supplémentaires permettraient de mieux distinguer la voix source de la voix cible.

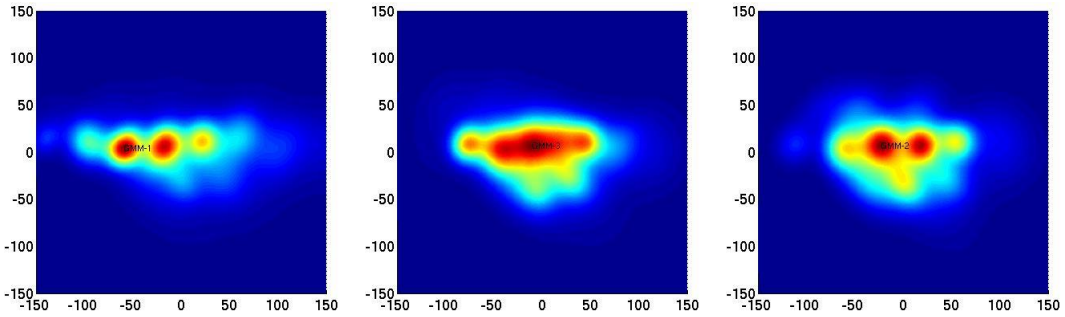

Fig. 1: Projection des densités de probabilité du GMM source 1(a), du GMM adapté 1(b), et du GMM cible 1(c). La projection est réalisée sur l'espace moyen des données source et cible. On peut noter le déplacement des densités du GMM source vers celles du GMM cible.

Tab. 3: Comparaison des valeurs de log-vraisemblance pour les GMM $\mathcal{M}_{\mathbf{X}}$, $\mathcal{M}_{\mathbf{Y}}$ et $\widehat{\mathcal{M}_{\mathbf{X}}}$ sur les ensembles de données de validation transformés.

	$\mathbf{X}'$	$\mathbf{Y}'$	$\mathbf{Z}'$
$\mathcal{M}_{\mathbf{X}}$	-18.60 ± 0.15	-19.80 ± 0.17	-19.34 ± 0.15
$\mathcal{M}_{\mathbf{Y}}$	-19.00 ± 0.16	-18.80 ± 0.17	-18.82 ± 0.16
$\widehat{\mathcal{M}_{\mathbf{X}}}$	-19.13 ± 0.16	-19.27 ± 0.16	-18.66 ± 0.15

6. Conclusions

Une méthodologie permettant de répondre au problème de transformation de la prosodie pour des corpus non-parallèles est présentée. La durée et le F_0, au niveau de la syllabe, sont représentés par un vecteur de dimension fixe. Un GMM est appris sur la durée et le F_0 du locuteur source. Nous proposons ensuite d'appliquer une approche MLLR pour adapter les paramètres du GMM source aux données du locuteur cible. Ensuite, un modèle permettant de transformer linéairement un vecteur prosodique est présenté. En utilisant ce modèle, la somme des centroïdes adaptés est pondérée par la probabilité a posteriori des vecteurs sources.

Le protocole expérimental repose essentiellement sur une validation croisée entre les modèles et les jeux de données. Une comparaison exhaustive entre les données et les modèles montre que le GMM adapté modélise efficacement les données cibles, et la fonction de transformation produit des données aussi vraisemblables pour le modèle cible que les données cibles elles-mêmes.

La méthodologie d'évaluation proposée est entièrement non parallèle. D'autres expériences sont planifiées pour confronter cette méthode aux méthodes classiques de transformation du F_0, dans un cadre parallèle. Compte-tenu de son influence, l'énergie pourrait également être intégré au vecteur de données.

Références

[1] J.A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, International Computer Science Institute, 1998.

[2] T. Ceyssens, W. Verhelst, and P. Wambacq. On the construction of a pitch conversion system. In *Proc. of EUSIPCO*, pages 1301–1304, 2002.

[3] D.T. Chappell and J.H.L. Hansen. Speaker-specific pitch contour modeling and modification. In *Proc. of ICASSP*, volume 2, pages 885–888, 1998.

[4] Laure Charonnat, Gaëlle Vidal, and Olivier Boeffard. Automatic phone segmentation of expressive speech. In *LREC*, 2008.

[5] A de Cheveigne and H Kawahara. Yin, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am*, 111 :1917–1930, 2002.

[6] M.J.F. Gales and P.C. Woodland. Mean and variance adaptation within the mllr framework. *Computer Speech & Language*, 10 :249–264, 1996.

[7] B. Gillett and S. King. Transforming f0 contours. In *Proc. of Eurospeech Conference*, pages 1713–1716, 2003.

[8] E.E. Helander and J. Nurminen. A novel method for prosody prediction in voice conversion. In *Proc. of ICASSP*, volume 4, pages 509–512, 2007.

[9] Z. Inanoglu. Transforming pitch in a voice conversion framework. Technical report, St. Edmond's College, Univ. of Cambridge, July 2003.

[10] L. Lamel, J.-L. Gauvain, and M. Eskenazi. Bref, a large vocabulary spoken corpus for french. In *European Conference on Speech Communication and Technology*, pages 505–508, 24-26 September 1991.

[11] P. Mertens. Automatic recognition of intonation in french and dutch. In *Proc. of Eurospeech conference*, pages 46–50, 1989.

[12] U.D. Reichel. Data-driven extraction of intonation contour classes. In *Proc. of the 6th ISCA SSW*, pages 240–245, 2007.

[13] Y. Stylianou, O. Cappé, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Tr. on Speech and Audio Processing*, 6 :131–142, 1998.

[14] Y. Yamashita, T. Ishida, and K Shimadera. A Stochastic F0 Contour Model Based on Clustering and a Probabilistic Measure. *IEICE Tr. on Information and Systems*, E86-D(3) :543–549, 2003.