Évaluation objective de plongements pour la synthèse de parole guidée par réseaux de neurones

Antoine Perquin[1] Gwénolé Lecorvé[1] Damien Lolive[1] Laurent Amsaleg[2]
(1) IRISA, 6 rue de Kerampont 22300 Lannion, France
(2) IRISA, 263 Avenue Général Leclerc, 35042 Rennes, France
{antoine.perquin, gwenole.lecorve, damien.lolive,
laurent.amsaleg}@irisa.fr

RÉSUMÉ

L'évaluation de plongements issus de réseaux de neurones est un procédé complexe. La qualité des plongements est liée à la tâche spécifique pour laquelle ils ont été entraînés et l'évaluation de cette tâche peut être un procédé long et onéreux s'il y a besoin d'annotateurs humains. Il peut donc être préférable d'estimer leur qualité grâce à des mesures objectives rapides et reproductibles sur des tâches annexes. Cet article propose une méthode générique pour estimer la qualité d'un plongement. Appliquée à la synthèse de parole par sélection d'unités guidée par réseaux de neurones, cette méthode permet de comparer deux systèmes distincts.

ABSTRACT

Objective evaluation of embeddings for speech synthesis guided by neural networks.

The evaluation of embeddings extracted from neural networks is complex. The quality of embeddings is relative to the task it was trained for and the evaluation of this task may be a lengthy and costly process if human annotators are involved. Thus, it may be useful to estimate their quality using fast and reproducible objective measures on auxiliary tasks. This paper introduces a generic method to estimate the quality of an embedding. This method is applied to speech synthesis based on unit selection guided by neural networks and allows to compare two systems.

MOTS-CLÉS : Plongements, Évaluation objective, Synthèse de parole.

KEYWORDS: Embeddings, Objective evaluation, Speech synthesis.

1 Introduction

La capacité de plongement des réseaux de neurones est souvent utilisée pour obtenir une représentation alternative de données. Extraire la sortie d'une couche cachée d'un réseau de neurones permet d'obtenir une représentation du vecteur d'entrée influencée par la tâche d'entraînement. La qualité d'un plongement est relative à une tâche donnée et l'évaluation de cette tâche peut être un procédé long et onéreux s'il y a besoin d'annotateurs humains.

En particulier, en synthèse de parole, l'évaluation d'un système est effectuée à l'aide de tests d'écoutes subjectifs. Afin d'obtenir des résultats rapidement et d'augmenter la reproductibilité des expériences, différentes mesures objectives existent. Cependant, ces mesures ne sont en général pas corrélées avec les résultats des tests d'écoutes et ne mesurent pas directement la qualité des plongements.

Cet article est un travail exploratoire visant à déterminer la qualité d'un plongement dans le cadre de la synthèse de parole. Nous proposons une méthode générique consistant à comparer un plongement dont on cherche à évaluer la qualité avec ceux issus d'entraînements volontairement défavorables en guise de références basses. La comparaison visuelle de ces plongements permet d'obtenir des critères distinctifs. La mise au point de mesures objectives correspondant à ces critères permet alors de comparer des plongements quelconques. L'application de cette méthode à la synthèse de parole indique qu'un plongement de qualité possède une structure par groupe de phonèmes et que la répartition de ces groupes est informée acoustiquement.

La suite de l'article suit le déroulement suivant : la section 2 présente diverses utilisations et méthodes d'évaluation de plongements, la section 3 présente une méthode générique d'estimation de leur qualité et les plongements utilisés. La section 4 présente l'observation visuelle de ces plongements et la section 5 l'élaboration puis l'utilisation de mesures objectives.

2 Travaux liés

La capacité de plongement des réseaux de neurones est utilisée en traitement automatique des langues pour obtenir des plongements de mots (Mikolov *et al.*, 2013). Il s'agit de représenter un mot avec un vecteur de petite dimension (relativement à la taille du vocabulaire) reflétant son contexte. La qualité des plongements peut être évaluée de manière intrinsèque ou extrinsèque (Schnabel *et al.*, 2015). Pour une évaluation extrinsèque, leur qualité est évaluée relativement à une tâche donnée (ex : reconnaissance d'entité nommée (Pennington *et al.*, 2014)) en les utilisant comme attributs d'entrées d'algorithmes d'apprentissage automatique. La qualité du modèle est alors conditionnée par la qualité des plongements utilisés. Pour une évaluation intrinsèque, les relations sémantiques entre mots issues des plongements peuvent être comparés à celles annotées par des humains (Mikolov *et al.*, 2013).

En synthèse de parole, les plongements sont utilisés par les méthodes par sélection d'unités guidée par réseaux de neurones. La méthode par sélection d'unités classique consiste à concaténer des unités de paroles pré-enregistrées afin d'obtenir un signal correspondant à un texte donné (Hunt & Black, 1996). La séquence d'unités à concaténer est choisie au sein d'une base de données comme la séquence qui minimise la somme de deux coûts. Le premier, coût de sélection, indique à quel point une unité dans la base de données est similaire à celle à synthétiser. Ce coût est habituellement défini par des experts linguistes. Le deuxième, coût de concaténation, indique à quel point deux unités consécutives dans une séquence d'unités se concatènent bien. Afin de diminuer la quantité d'expertise linguistique nécessaire pour élaborer des systèmes de synthèse par sélection d'unités, le coût de sélection peut être remplacé par les prédictions d'un réseau de neurones (Merritt *et al.*, 2016), ou la distance euclidienne dans un espace de plongement défini par la couche cachée d'un réseau de neurones (Wan *et al.*, 2017; Perquin *et al.*, 2018). On parle alors de synthèse de parole par sélection d'unités guidée par réseaux de neurones. L'objectif de cet article est la mise au point de mesures objectives de la qualité des plongements servant à guider cette sélection d'unités.

La qualité des plongements pour la synthèse de parole est habituellement évaluée de manière extrinsèque, par des tests d'écoutes. La qualité d'un système seul peut être évaluée à partir de notes arbitraires (Union, 1996) ou par comparaison avec d'autres systèmes (Union, 2003). Cependant, ces tests demandent de nombreux participants pour contre-balancer l'aspect subjectif et être fiables. Afin d'obtenir des indices de qualité de manière rapide et reproductible, différentes mesures objectives existent. Par exemple, la distorsion mel-cepstrale (MCD) permet de mesurer une différence acoustique

entre un signal produit par le système et un signal de référence (Kubichek, 1993). Cette mesure est assimilable à une erreur de reconstruction par le réseau de neurones. Cependant, ces mesures objectives ne sont pas directement des indicateurs extrinsèques de la qualité d'un plongement car elles sont habituellement utilisées pour mesurer la qualité des prédictions du réseau dont sont issus les plongements. L'une des mesures proposée par cette article consiste à les utiliser de manière réellement extrinsèque, en évaluant les prédictions d'un modèle entraîné sur les plongements.

Ce travail présente une méthode générique pour mettre au point des mesures objectives de la qualité d'un plongement. Appliquée à la synthèse de parole guidée par réseaux de neurones, nous proposons différentes mesures de qualité extrinsèque pour des plongements de phones.

3 Protocole expérimental

Ce travail est le fruit d'une réflexion sur la qualité d'un plongement et les mesures associées. Nous proposons d'identifier des critères de qualité en comparant un plongement générique avec un plongement jugé mauvais par construction. Cette section présente la méthode employée et les plongements envisagés.

3.1 Méthode

La méthode présentée ici est issue de deux constats. Premièrement, il est plus simple de mettre au point une mesure objective distinguant des plongements de qualités très différentes que des plongements de qualités similaires. La méthode propose donc de se concentrer sur l'obtention de mesures permettant de distinguer un plongement quelconque d'un plongement de mauvaise qualité. On peut ensuite vérifier que ces mesures permettent aussi de distinguer des plongements de qualité quelconque. Sous l'hypothèse que la qualité d'un plongement peut être influencée par l'apprentissage du réseau de neurones correspondant, la méthode propose d'entraîner des plongements dans des conditions défavorables afin d'obtenir les plongements supposés de mauvaise qualité. Deuxièmement, il n'est pas toujours intuitif de trouver un critère de qualité d'un plongement. La méthode propose donc de visualiser des espaces de plongements à l'aide d'une méthode de réduction de la dimensionalité. La comparaison visuelle des espaces peut permettre de déduire des critères distinctifs qui serviront de base à la définition de mesures objectives de la qualité d'un plongement.

La méthode proposée pour mettre au point des mesures de la qualité d'un plongement est la suivante :

1. Entraînement de plongements, éventuellement issus de modèles différents, dont on souhaite estimer la qualité ;
2. Entraînement de plongements dans des conditions sous-optimales, ils sont jugés mauvais par construction ;
3. Comparaison visuelle des plongements pour obtenir une intuition de critères distinctifs ;
4. Mise au point de mesures correspondant à ces critères distinctifs ;
5. Comparaison des plongements issus de l'étape 1 avec ceux issus de l'étape 2 afin de vérifier que les mesures objectives correspondent aux critères distinctifs identifiés ;
6. Comparaison des plongements issus de l'étape 1 entre eux pour vérifier que les mesures sont distinctives dans le cas général.

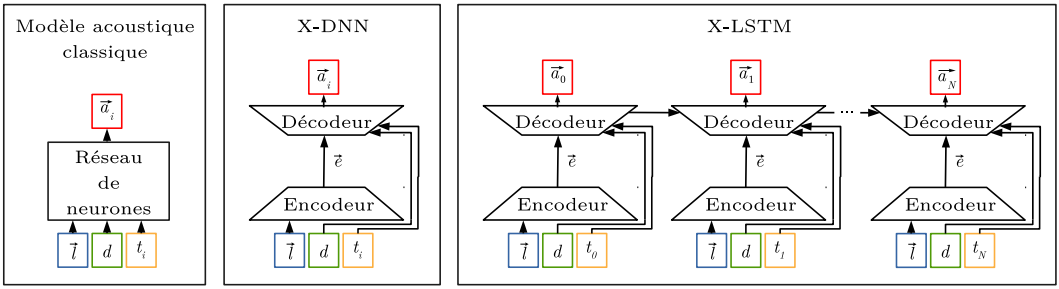

FIGURE 1 – Schématisation d'un modèle acoustique classique et des modèles proposés pour l'extraction de plongements.

Cette étude se concentre sur l'étude de plongements de phones dans le cadre de la synthèse de parole guidée par réseaux de neurones. Les plongements sont obtenus à partir d'un modèle acoustique (schématisé sur la figure 1). Pour un phone de durée d décrit linguistiquement par $\vec{l}$ (identité du phone, de ses voisins, position dans le mot, etc.), pour une trame de position t_i décrite acoustiquement par $\vec{a_i}$ (mel-cepstrum, bande d'apériodicité et fréquence fondamentale), le réseau tente de prédire $\vec{a_i}$ en fonction de $\vec{l}$, d et t_i. Afin d'obtenir des plongements de phones plutôt que de trames, d et t_i ne sont fournies qu'après la couche cachée servant de projection.

Sous l'hypothèse que la qualité d'un plongement peut être influencée par l'apprentissage du modèle correspondant, les plongements supposés mauvais seront entraînés dans des conditions volontairement défavorables. Ici, il s'agira d'un sous-apprentissage et d'un sur-apprentissage. Ces mauvais plongements sont choisis car ils peuvent être appris rapidement (peu de données et/ou d'époques d'apprentissage). De plus, en comparaison avec un espace de plongement complètement aléatoire, ils devraient représenter des références basses plus pertinentes. La comparaison visuelle sera effectuée à l'aide d'une projection des plongements par Analyse en Composantes Principales (ACP). Il est important de remarquer que d'autres méthodes de réduction de la dimensionalité pourrait être utilisées. Alors, la visualisation obtenue serait différente et les critères distinctifs pouvant être déduits le seraient aussi.

3.2 Modèles

Le premier modèle proposé pour l'extraction de plongements de phones, X-DNN, peut être divisé en deux parties. L'encodeur est composé de 5 couches de dimensions 1024 à 64. Il permet de plonger la description linguistique $\vec{l}$ d'un phone dans un espace vectoriel de dimension 64. Le décodeur est composé de 4 couches cachées de dimensions 128 à 1024 avec une couche finale de dimension 199. À partir d'un vecteur de plongement $\vec{e}$ de phone, de la durée d du phone et de la position t_i d'une trame au sein de ce phone, le décodeur permet de prédire les attributs acoustiques $\vec{a_i}$ associés à la trame. Puisque la fonction d'activation de chaque couche cachée est une tangente hyperbolique, les plongements prennent des valeurs dans l'intervalle $[-1, 1]$.

Le modèle X-LSTM reprend l'architecture de X-DNN en remplaçant la première couche cachée du décodeur par une couche LSTM afin de modéliser les dépendances temporelles au sein d'un phone. En réalité, la présence de cette couche oblige à placer toutes les trames du phone dans un même *batch*, ce qui entraîne une prédiction lissée des $\vec{a_i}$. Malgré des prédictions acoustiques imprécises, les

FIGURE 2 – Visualisation des plongements issus de ref-DNN

FIGURE 3 – Visualisation des plongements issus de sous-DNN

FIGURE 4 – Visualisation des plongements issus de sur-DNN

plongements résultants de X-LSTM permettent une synthèse par sélection d'unités satisfaisante (cf. Tableau 1). Plus d'informations sur les modèles, les attributs utilisés et les performances associées sont disponibles dans (Perquin *et al.*, 2018).

Le jeu de données d'entraînement contient une dizaine d'heures de parole pour un locuteur français masculin professionnel (jeu de données non publié). Cela correspond à 3 300 énoncés soit environ 390 000 phones. La parole est expressive (narration, dialogues joués) et les phrases sont complexes (longues, registre soutenu). Le jeu de donnés est divisé en un sous-ensemble d'entraînement (90%), de test (5%) et de développement (5%).

Les modèles X-DNN et X-LSTM sont entraînés sur la totalité du jeu d'entraînement disponible afin d'obtenir les modèles de références ref-DNN et ref-LSTM (meilleurs modèles sur 100 époques). Les plongements sous-optimaux sont obtenus en entraînant l'architecture X-DNN dans deux cas de mauvais apprentissage. Le modèle sous-DNN est obtenu en entraînant l'architecture sur 256 trames pendant une seule époque. Il s'agit d'un cas de sous-apprentissage. Le modèle sur-DNN est obtenu en entraînant l'architecture sur 256 trames pendant 50 époques. Il s'agit d'un cas de sur-apprentissage. Le nombre de trames d'entraînement pour les modèles sous-DNN et sur-DNN a été choisi pour correspondre à la taille d'un *batch* pour le modèle ref-DNN. Chaque trame correspond à un phone différent, ces modèles ne sont donc appris que sur 256 phones choisis aléatoirement.

4 Intuition visuelle

La méthode proposée consiste à définir des mesures de qualité en comparant un plongement donné à un plongement jugé mauvais par construction. Cependant, avant de mettre au point ces mesures, il est nécessaire d'obtenir une intuition sur les critères qui permettent de distinguer les plongements considérés. Nous proposons d'obtenir cette intuition en observant visuellement les plongements, ici par ACP. Chaque point correspond au plongement d'un phone, la couleur de ces points indique le phonème associé.

La figure 2 représente la visualisation par ACP du plongement des phones par le modèle ref-DNN. Graphiquement, il semble que les points de même couleur sont regroupés. Cela signifierait que le plongement de référence permet de grouper les phones associés au même phonème. De plus, dans cet

espace projeté, il semble que les plongements associés aux phones de /p/ et /t/ d'une part et /e/ et /ɛ/ d'autre part sont proches, tandis que ceux des /p/ et des /e/ sont éloignés. Cela indiquerait que l'espace de plongement de référence tient compte d'une similarité acoustique dans la répartition des groupes de phones.

La figure 3 représente la visualisation par ACP du plongement des phones par le modèle sous-DNN. De manière similaire, il semble qu'un plongement sous-appris regroupe les phones par phonème et que la répartition de ces groupes soient informée par une similarité acoustique. En revanche, il semblerait que la distinction entre groupes de phones soit moins claire dans l'espace de plongement sous-appris que dans celui de référence.

La figure 4 représente la visualisation par ACP du plongement des phones par le modèle sur-DNN. Visuellement, il semble impossible de distinguer une structure particulière. En particulier, aucune répartition en groupe de phones n'est remarquable.

La visualisation effectuée n'est en rien une preuve de la qualité d'un plongement ou un reflet exact de sa structure. Cependant, en comparant visuellement le plongement de référence avec les plongements sous-optimaux, deux critères de qualité émergent : une structure de groupe par phonème, une répartition de ces groupes informée acoustiquement. Ces critères sont au final assez intuitifs, mais une méthode de visualisation différente pourrait peut-être permettre de déduire d'autres critères distinctifs.

5 Mesures objectives

Cette section présente la mise au point de mesures correspondant aux critères distinctifs issus de l'observation visuelle puis leur application.

5.1 Définition des mesures

Le premier critère de qualité proposé est celui de la structure par groupe de phonèmes. Pour évaluer ce critère, on propose de s'intéresser à la notion de plus proches voisins. Soit e le plongement d'un phone quelconque, e_i pour $i \in [1, 100]$ les 100 plus proches voisins de e. Si la structure de l'espace de plongement regroupe les phones par groupe de phonèmes, les plus proches voisins du phone plongé doivent correspondre au même phonème. Ce critère peut être mesuré de deux manières complémentaires :

- Comparaison de la classe majoritaire parmi les e_i avec la classe de e, en considérant que la classe d'un phone est le phonème associé. On s'intéresse alors à la précision d'une classification par plus proches voisins.
- Mesure du pourcentage des e_i partageant la même classe que e. On s'intéresse alors à la pureté du voisinage de e.

Le second critère de qualité suggéré par la visualisation des plongements est la répartition des groupes de phonèmes en fonction d'une similarité acoustique. Afin d'éviter l'utilisation d'expertise pour la définition de cette similarité acoustique entre phonèmes, on propose ici de s'intéresser à la mesure du potentiel de prédiction acoustique d'un plongement. Pour une trame d'un phone, un modèle de régression linéaire est entraîné à prédire les coefficients acoustiques $\vec{a_i}$ de la trame en fonction du plongement $\vec{e}$ du phone, de sa durée d et la position t_i de la trame. Ce modèle est entraîné sur le jeu

	Linguistique	ref-DNN	sous-DNN	sur-DNN	Aléatoire	ref-LSTM
Classification (précision)	**0.972**	0.952	0.893	0.882	0	0.930
Pureté (pourcentage)	**92.9** a†	92.2 a†	84.3	81.5	4.53	89.6
MCD (dB)	6.02	**5.84**	6.66 b†	6.70 b†	8.21	6.20
Test d'écoute (/10) (Perquin *et al.*, 2018)		**7.0**				6.4

TABLE 1 – Mesures objectives pour chaque plongement. † Au sein d'une ligne, suivies de la même lettre, les différences de mesures ne sont pas significatives (0.05)

d'entraînement grâce à la méthode des moindres carrés. La qualité d'un plongement est alors mesurée de manière extrinsèque en calculant la MCD (Distortion Mel Ceptrale) moyenne entre les valeurs prédites et réelles.

5.2 Mesures expérimentales

Les mesures définies dans la section 5.1 sont appliquées aux modèles ref-DNN, sous-DNN et sur-DNN. Pour comparaison, elles sont aussi appliquées sur l'espace des descripteurs linguistiques $\vec{l}$, un espace de plongement aléatoire (vecteurs de dimension 64 aléatoirement uniformes sur $[-1, 1]$) et l'espace de plongement défini par ref-LSTM. Les résultats de ces mesures sont rapportés dans le tableau 1.

Par comparaison avec l'espace linguistique, les plongements issus de ref-DNN obtiennent des mesures de classification et de pureté similaire, mais une meilleure mesure acoustique. Ainsi, un plongement correctement entraîné semble conserver la structure de l'espace d'origine, tout en offrant une meilleure capacité de prédiction pour la tâche d'entraînement. En revanche, pour les plongements issus de sous-DNN et sur-DNN, les mesures liées à la conservation linguistique et à la prédiction acoustique sont toutes plus faibles que pour l'espace linguistique. Alors, un plongement mal entraîné semble perdre une partie de la structure de l'espace linguistique et perd même une partie du potentiel de prédiction vis-à-vis de la tâche d'entraînement.

Pour ref-LSTM, les résultats sont légèrement en dessous de ceux de ref-DNN pour toutes les mesures. Cela semble indiquer que les plongements issus de ref-LSTM sont moins bons que ceux issus de ref-DNN. Ces résultats sont cohérents avec les tests d'écoutes effectués dans (Perquin *et al.*, 2018).

6 Conclusion

Une méthode est proposée afin d'identifier des critères de qualité d'un plongement, avant de mettre au point des mesures objectives associées. Cette méthode est applicable à tous les plongements, indépendamment de la tâche d'entraînement. Dans le cas de la synthèse de parole, un plongement doit grouper les phones par phonèmes, et ces groupes doivent être répartis selon une similarité acoustique. Les mesures objectives proposées évaluent la qualité du plongement de manière extrinsèque via une classification par plus proches voisins et un modèle de régression linéaire. Cependant, des méthodes de régression plus complexes pourrait permettre de mieux distinguer les plongements. De plus, des méthodes de visualisation autres que l'ACP permettent de mieux conserver les relations de voisinages, ce qui pourrait permettre la découverte de nouveaux critères distinctifs.

Références

HUNT A. J. & BLACK A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

KUBICHEK R. (1993). Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of the Pacific Rim Conference on Communications, Computers and Signal Processing (CCSP)*.

MERRITT T., CLARK R. A., WU Z., YAMAGISHI J. & KING S. (2016). Deep neural network-guided unit selection synthesis. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the Annual Conference on Advances in Neural Information Processing Systems (NIPS)*.

PENNINGTON J., SOCHER R. & MANNING C. (2014). Glove : Global vectors for word representation. In *Proceedings of the Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

PERQUIN A., LECORVÉ G., LOLIVE D. & AMSALEG L. (2018). Phone-level embeddings for unit selection speech synthesis. In *Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP)*.

SCHNABEL T., LABUTOV I., MIMNO D. & JOACHIMS T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

UNION I. T. (1996). Methods for subjective determination of transmission quality. *ITUT Recommendation*.

UNION I. T. (2003). Method for the subjective assessment of intermediate quality level of coding systems. *ITUT Recommendation*.

WAN V., AGIOMYRGIANNAKIS Y., SILEN H. & VIT J. (2017). Google's next-generation real-time unit-selection synthesizer using sequence-to-sequence lstm-based autoencoders. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*.