Visualisation intuitive des variations prosodiques pour l'entraînement à la prise de parole en public

Intuitive Visualisation of Prosodic Variations for Public Speaking Coaching

Nesrine Fourati, Mathieu Chollet, Chloé Clavel

English Abstract—The tracking of temporal changes in a speaker's vocal behavior can greatly enhance the development of feedback-based public speaking coaching. Related studies tend to focus on the overall score of the performance and/or detailed feedback related to a short sequence of speech. In this paper, we present our visual feedback approach aiming to offer a good trade-off between 1) capturing the details of the temporal changes in the speaker's performance and 2) offering intuitive and hopefully reliable feedback.

1 Introduction et Etat de l'art

Les compétences sociales, en particulier les compétences de Prise de Parole en Public (PPP), sont devenues cruciales dans les contextes professionnels modernes. Avec l'utilisation croissante des vidéo-conférences suite à la pandémie du Covid-19, la communication professionnelle se transforme, et maîtriser sa compétence orale devient crucial. L'expressivité vocale constitue l'une des principales modalités de communication dans une PPP [4], [13]. La maîtrise de la modulation de la prosodie est cruciale pour la bonne maîtrise de la PPP [9], et ce fort intérêt a stimulé la création d'une large gamme de systèmes de formation à la PPP assistés par ordinateur [5], [9], [12]–[14]. L'un des principaux défis pour le développement d'un système de formation aux compétences sociales pour la PPP consiste à offrir un feedback intuitif, exploitable et fiable [3], [8]. Les systèmes de coaching vocal existants ont tendance à se focaliser sur la pratique de segments vocaux très courts, généralement d'une à deux phrases [11], [13]. Cette approche suppose qu'en travaillant sur la pratique des compétences vocales sur des segments courts, l'utilisateur sera en mesure de les exploiter au cours d'une prise de parole complète. [13]. D'autres systèmes de coaching de PPP basés sur le feedback se concentrent sur des discours plus longs, suivant principalement le mantra « Vue d'ensemble d'abord, zoom et filtrage, puis détails à la demande » [14], un paradigme d'interaction dans lequel les utilisateurs font des allers-retours entre une vue d'ensemble globale et une vue des détails locaux de leur discours [5], [14]. Web-Pitcher [10] est un autre système d'entraînement à la PPP, il vise à pratiquer des compétences de modulation vocale dans de longs discours. Le système Web-Pitcher met à jour le feedback visuel toutes les 5 secondes [9].

Dans l'ensemble, nous constatons que peu d'attention a été accordée à la visualisation des changements temporels de l'expressivité vocale de l'orateur pour le développement d'un système d'entraînement à la PPP. Nous présentons une approche de visualisation intuitive du feedback donné sur la performance vocale lors d'une PPP, où l'objectif est de trouver un bon compromis entre un vue globale et un retour détaillé sur le comportement vocal pendant une PPP complète. Basée sur une technique d'interpolation polynomiale et une approche axée sur les données, notre méthode est capable de résumer les changements temporels dans le comportement vocal lors d'une PPP, permettant ainsi à l'utilisateur de s'entraîner pour améliorer et maîtriser sa compétence orale dans une PPP.

2 Approche proposée

Ce travail vise à concevoir un feedback intuitif, exploitable et fiable sur l'expressivité vocale dans une PPP. La question de recherche principale que nous abordons ici est la suivante: Comment fournir un feedback intuitif permettant de visualiser les changements clés de la variation prosodique, permettant ainsi d'auto-évaluer la variation de sa voix?

Dans un premier temps, nous nous concentrons sur une seule caractéristique vocale; le débit de parole qui a été identifiée comme une des caractéristiques clés de la modulation de la voix [9], [13]. Nous

- *Nesrine Fourati: CGI, Centre d'innovation digitale*
  E-mail: nesrine.fourati@cgi.com
- *Mathieu Chollet:School of Computing Science, University of Glasgow, United Kingdom*
  E-mail: Mathieu.Chollet@glasgow.ac.uk
- *Chloé Clavel: LTCI, Télécom Paris, Institut Polytechnique de Paris, France*
  E-mail: chloe.clavel@telecom-paris.fr

souhaitons inclure dans nos futurs travaux d'autres caractéristiques vocales tels que les pauses, la tonalité et l'intensité de la voix ainsi que d'autres caractéristiques du comportement non-verbal. Nous utilisons la base de données "Ma thèse en 180 secondes" [1], qui comprend 245 vidéos (135F, 110M), d'une durée de 3 minutes chacune, issues d'un concours français scientifique.

Extraction automatique du débit de parole:

Le débit de parole est souvent quantifié en termes de mots ou de syllabes par minute [6]. Ici, nous le définissons comme le nombre de syllabes par seconde dans une fenêtre de temps donnée. Nous écartons les segments non voisés (i.e. les silences) pour mesurer le taux d'articulation dans les intervalles voisés uniquement. Ainsi nous nous focalisons sur le rythme auquel la parole est énoncée [6].

De nombreuses solutions open-source sont disponibles pour l'analyse acoustiques: nous utilisons la librairie Parselmouth qui permet d'accéder aux fonctionnalités de PRAAT [2] (v 0.4.2), introduit par Jadoul et al. [7]. Nous avons adapté et utilisé un script Praat afin de détecter les syllables ainsi que les segments non voisés[1].

Afin d'obtenir un premier feedback visuel intuitif sur le débit de parole, nous optons pour une approche de visualisation basée sur les courbes. Ce type de visualisation a déjà été utilisé pour le développement des systèmes de feedback pour l'entraînement à la PPP [3], [5], [9], [13]. Les valeurs de réfèrences pour le débit de parole ont été définies en se basant sur les données du corpus comme suit: la valeur minimale absolue ainsi que trois quartiles correspondants aux données du 25e, 50e, 75e et 99e percentiles des données permettant de donner respectivement quatre références pour un débit de parole Très Lent, Lent, Normal, Élevé et Très Élevé.

Lissage des variations locales avec une fenêtre glissante:

Étant donné que nous souhaitons trouver le meilleur compromis entre la visualisation des variations vocales globales et locales, nous nous orientons en premier lieu vers les techniques de lissage de courbe afin de réduire les irrégularités et singularités des changements locaux d'une performance vocale. Pour cela nous proposons en premier lieu de lisser les mesures originales du débit de parole avec un algorithme de fenêtre glissante qui s'inspire de l'algorithme de Savitzky-Golay; le débit de parole est calculé de manière récurrente sur une fenêtre de temps glissante d'une durée et un pas déterminé.

Dans la suite de ce papier, nous appelons 1) la fenêtre de temps, 2) la fenêtre de temps glissante, et 3) la durée de la fenêtre de temps glissante respectivement FT, FTG et DFTG.

1. https://github.com/Shahabks/myprosody/blob/master/myprosody/dataset/essen/myspsolution.praat

Les mesures originales du débit de parole sont affichées par une courbe noire dans le premier graphe de la figure 1. Les courbes noires dans les graphes 2, 3, 4 de la Figure 1 montrent l'évolution progressive du lissage obtenu par application de fenêtres glissantes avec des largeurs progressivement plus grandes; les fréquences plus élevées du signal étant progressivement plus filtrées par la FT plus grande agissant comme un filtre passe-bas. Nous fixons le pas des FTG successives à 1 seconde afin de garantir le maximum de chevauchement entre deux fenêtres de temps successives. Cependant, le choix de la durée de la fenêtre de temps glissante reste à déterminer (plus la fenêtre est large, plus la courbe est lissée). Une courte DFTG conduit à une courbe bruyante car de nombreux détails sont visualisés (voir graphe 2 de la figure 1; DFTG=5sec). Une grande DFTG conduit à un fort lissage des changements temporels avec une perte potentielle de moments clés significatifs de variation vocale. Une DFTG adéquate et un court pas permettent de lisser la courbe des données originales tout en capturant les changements temporels majeurs dans le comportement vocal.

Lissage des variations locales avec un ajustement de la courbe:

En plus de l'approche de lissage de courbe par fenêtre glissante décrite dans le paragraphe précédent, les approches d'ajustement de courbe (« Curve fitting ») peuvent également être utilisées pour lisser les variations locales d'une série de données. La courbe rouge dans le premier graphe de la figure 1 illustre le résultat de l'interpolation polynomiale ajustée aux mesures originales du débit de parole. Nous constatons que, même avec un degré assez haut (égale à 30), le coefficient de détermination R^2 (qui permet de mesurer la qualité de l'ajustement de la courbe) obtenu est assez bas (e.g. R^2 = 0.3, voir graphe1 de la figure 1). Ceci montre que le résultat de la courbe ajustée aux données brutes en appliquant une simple interpolation polynomiale ne peut pas être considéré comme suffisamment représentatif des principaux changements locaux.

Combinaison des deux approches:

Nous proposons de combiner les deux approches discutées ci-dessus afin de trouver la meilleure approximation permettant de résumer les principaux variations de la modulation vocale. Nous proposons d'ajuster un modèle d'interpolation polynomiale sur le résultat donné par l'algorithme de la fenêtre glissante. Comme nous visons à identifier la meilleure DFTG, nous augmentons les valeurs de la DFTG de 5sec à 20sec en augmentant progressivement la DFTG d'une seconde. Il en résulte 15 configurations. La dimension du polynôme est initialement fixée à 30. D'autres expériences seront menées dans nos futurs travaux pour explorer davantage l'ajustement du polynôme de dimension inférieure. Les courbes rouges dans le deuxième, troisième et

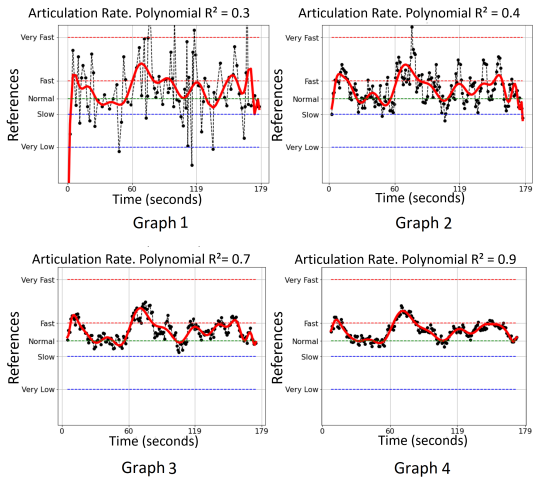

Fig. 1. La variation du débit de parole pour un échantillon de données extrait du corpus 3MTDataset: 1) les données originales du débit de parole (noir) et l'interpolation polynomiale correspondante (rouge), 2, 3) et 4) le résultat du lissage par la fenêtre glissante (noir) avec une DFTG = 5sec, 10sec et 15sec et l'interpolation polynomiale correspondante.

quatrième graphes de la figure 1 illustrent le résultat de l'interpolation polynomiale ajustée aux données obtenues par le lissage avec l'algorithme de la fenêtre glissante, avec des durées progressivement plus grandes.

Résultats:

L'évolution de l'erreur de l'ajustement de l'interpolation polynomiale (1 - R^2) est mesurée pour chacune des 15 configurations présentées ci-dessus, et ce pour la totalité des échantillons présents dans le corpus (environ 245 vidéos). Comme prévu, nous avons observé que l'erreur de l'interpolation polynomiale diminue progressivement jusqu'à obtenir approximativement une correspondance parfaite avec la plus large DFTG (c.-à-d. 20 secondes). Ce résultat est attendu car plus nous augmentons la DFTG, plus nous lissons les fortes variations locales. Par conséquence, plus on augmente la DFTG, plus il est plus facile pour un polynôme d'ajuster les données (voir figure 1).

Pour identifier la DFTG optimale, nous procédons à une analyse statistique, étant à la recherche de la DFTG qui offre un score d'ajustement stable pour le modèle d'interpolation (c'est-à-dire que l'erreur du modèle d'interpolation ne diminue plus de manière significative en augmentant la DFTG). Nous utilisons des tests post-hoc Pairwise Tukey-HSD, basés sur les scores d'ajustement obtenus pour tous les échantillons du corpus 3MT, pour augmenter la DFTG (de 5s à 20s). En se basant sur l'approche décrite ci-dessus, nous avons trouvé qu'à partir d'une durée égale à 14 - 15s, (14s pour α=0.01 et 0.001, et 15s pour α=0.05), l'erreur de l'ajustement de l'interpolation se stabilise; les changements obtenus en augmentant la DFTG ne passent pas le test de signification. En d'autres termes, nous avons constaté qu'une DFTG de 14-15s (avec un pas = 1 seconde) semble être appropriée pour résumer les principales variations du débit de parole pour l'ensemble de données 3MT.

Les courbes noire et rouge du graphe 4 de la figure 1 illustrent respectivement 1) la courbe résultante d'un lissage avec une DFTG égale à 15 secondes (avec un pas de 1sec) ainsi que 2) l'ajustement polynomial résultant. Nous pouvons observer que la courbe de l'interpolation polynomiale s'ajuste très bien avec le résultat du lissage obtenu par l'algorithme de la fenêtre glissante avec une DFTG = 15sec. Cette courbe semble donc être bien représentative des principaux changements locaux du débit de parole, permettant de fournir un feedback intuitif et fiable sur la variation temporelle du débit de parole sur une PPP complète.

Dans l'exemple de la courbe montrée dans le graphe 4 de la figure 1, le débit de parole élevé est d'abord observé pendant les toutes premières secondes, puis les valeurs oscillent légèrement autour de la référence intermédiaire d'un débit "Normal" jusqu'à 1 minute (60s), où le locuteur commence à accélérer son débit de parole, pour finalement terminer par une série de variations entre les bornes de références "Lent" et "Rapide". Ces variations seraient difficiles à percevoir à partir des données brutes du débit de parole.

3 Conclusion

L'objectif principal de ce travail est de répondre à l'un des principaux défis de recherche en matière de feecback sur les compétences sociales: comment concevoir des approches intuitives et exploitables basées sur un feedback offrant un niveau de détail idéal (suffisamment de détails pour identifier les faibles ou fortes variations du comportement du locuteur, avec un minimum de détails susceptibles de submerger les utilisateurs).

Bien que nous nous focalisons ici sur une seule caractéristique vocale, nous estimons que notre approche de feedback visuel peut être généralisée à d'autres caractéristiques non verbales (cela fait partie de nos futurs travaux de recherche). Nous nous sommes également focalisé dans cet article sur un module de Feedback direct, mais nous travaillons actuellement sur des fonctionnalités de feedback indirect qui permettraient de comparer et de classer les performances de PPP de l'orateur en se basant sur des techniques de Machine Learning.

Nous comptons conduire d'autres expérimentations et analyses dans nos futurs travaux de recherche pour explorer; la dimension la plus appropriée du modèle polynomiale, d'autres techniques d'approximation, ainsi qu'une étude perceptuelle pour évaluer le feedback visuel.

REFERENCES

[1] Beatrice Biancardi, Mathieu Chollet, and Chloé Clavel. Introducing the 3MT_French Dataset to Investigate the Timing of Public Speaking Judgements. *PREPRINT (Version 1) available at Research Square*, 10 2022.

[2] P. Boersma and V. van Heuven. Speak and unSpeak with Praat. *Glot International*, 5(9-10):341–347, 2001.

[3] M. Chollet, S. Marsella, and S. Scherer. Training public speaking with virtual social interactions: effectiveness of real-time feedback and delayed feedback. *Journal on Multimodal User Interfaces*, 16(1):17–29, 3 2022.

[4] K. Fischer, O. Niebuhr, M. Alm, Abelin, E. Albertsen, and A. Asadi. Towards a Prosodic Visualization Tool for Language Learners. pages 269–272. International Speech Communication Association, 7 2022.

[5] M. Fung, Y. Jin, R. Zhao, and M. E. Hoque. ROC speak: Semi-Automated Personalized Feedback on Nonverbal Behavior from Recorded Videos. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15*, pages 1167–1178, New York, New York, USA, 2015. ACM Press.

[6] E. Jacewicz, R. A. Fox, C. O'Neill, and J. Salmons. Articulation rate across dialect, age, and gender. *Language Variation and Change*, 21(2):233–256, 7 2009.

[7] Y. Jadoul, B. Thompson, and B. de Boer. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71(2018):1–15, 2018.

[8] E. H. Mory. Feedback research revisited. In *Handbook of research on educational communications and technology*, pages 738–776. Routledge, 2013.

[9] O. Niebuhr. Computer-assisted prosody training: Improving public speakers' vocal charisma with the Web-Pitcher. *Revista da ABRALIN*, pages 1–29, 9 2021.

[10] O. Niebuhr, A. Brem, and S. Tegtmeier. Advancing research and practice in entrepreneurship through speech analysis – From descriptive rhetorical terms to phonetically informed acoustic charisma profiles. *Journal of Speech Sciences*, 6(1):3–26, 11 2017.

[11] R. S. Schaefer, L. J. Beijer, W. Seuskens, T. C. Rietveld, and M. Sadakata. Intuitive visualizations of pitch and loudness in speech. *Psychonomic Bulletin and Review*, 23(2):548–555, 2016.

[12] H. Tanaka, S. Sakti, G. Neubig, H. Negoro, H. Iwasaka, and S. Nakamura. Automated social skills training with audiovisual information. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2262–2265. IEEE, 8 2016.

[13] X. Wang, H. Zeng, Y. Wang, A. Wu, Z. Sun, X. Ma, and H. Qu. VoiceCoach: Interactive Evidence-based Training for Voice Modulation Skills in Public Speaking. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, New York, NY, USA, 4 2020. ACM.

[14] H. Zeng, X. Wang, A. Wu, Y. Wang, Q. Li, A. Endert, and H. Qu. EmoCo: Visual Analysis of Emotion Coherence in Presentation Videos. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2019.