Influence de la décision voisé/non-voisé dans l'évaluation comparative d'algorithmes d'estimation de F_0

François Signol, Jean-Sylvain Liénard, Claude Barras

LIMSI-CNRS – Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
Bâtiments 502bis et 508, Université Paris-Sud, 91400 Orsay Cedex, France
{francois.signol, jean-sylvain.lienard, claude.barras}@limsi.fr
http ://www.limsi.fr/Individu/{signol, lienard, barras}

ABSTRACT

This paper aims at pointing out the impact of the voiced/unvoiced decision on the results obtained by any pitch estimation algorithm. To be equitable, a comparative evaluation must provide the voided/unvoiced state in terms of overvoiced rate and undervoiced rate. The interaction between the detection and the estimation is usually underestimated in the pitch evaluation framework. Its consequence is a large bias in the comparative evaluation. This paper illustrates this phenomenon through the use of two comparative evaluation methodologies on 25 minutes of voiced speech. PEA could be improved not only from the F0 estimation part but also from the voiced/unvoiced decision part. This raises the problem of quantifing properly the voicing strength of a speech frame.
Keywords: F_0 estimation, voiced/unvoiced decision, comparative evaluation.

1. Introduction

Le problème d'estimation de la F_0 d'un signal monopitch est un problème ancien dont les principes sont donnés dans [6] et qui reste encore largement ouvert. Dans le domaine de la parole, il existe un grand nombre d'algorithmes d'estimation de pitch (AEP). L'évaluation des AEP est une étape indispensable et l'usage qui est fait des résultats obtenus conditionne la complexité du processus d'évaluation. L'évaluation la plus simple est celle qui consiste à donner des indicateurs de performance objectifs pour un AEP donné à un instant donné. Le processus se complique lorsque l'objectif est de quantifier les progrès apportés au même AEP dans le temps. Le processus d'évaluation le plus compliqué est certainement celui dont le but est de fournir une comparaison des performances d'AEP différents. Il impose en effet d'être *équitable* alors que les AEP testés peuvent être très hétérogènes dans leur but et leur fonctionnement. Une évaluation comparative doit au minimum uniformiser les corpus qui sont de même nature de parole (lue, spontanée, etc.), l'intervalle de recherche de la F_0 et les instants temporels d'estimation. Un léger décalage peut aboutir à plusieurs pour-cents d'erreur en plus (ou en moins). Ces points peuvent paraître évidents mais leur influence sur les résultats est souvent sous-estimée. Dans la mesure du possible, il est aussi préférable d'homogénéiser la durée des trames et de prendre en compte les éventuels post-traitements correctifs (prise en compte du contexte temporel).

De nombreuses évaluations de la qualité des AEP existent. C'est dans le domaine musical que l'évaluation est la plus dynamique sous l'impulsion de MIREX [1]. Cette campagne d'évaluation annuelle propose de nombreux travaux concernant les meilleures méthodologies à suivre pour évaluer des AEP multipitch. L'analyse de cette littérature met en évidence une analogie entre le problème de détection de l'étendue temporelle d'une note dans une tâche de transcription automatique de piano et celui de la décision voisé/non-voisé (VnV) en parole [5]. De telles campagnes d'évaluation n'existent pas encore pour la parole. Il existe néanmoins des travaux concernant l'évaluation comparative d'AEP [3][10][9]. Dans ce domaine, l'estimation de F_0 sur des signaux monolocuteur relativement neutres d'expressivité et enregistrés en condition de laboratoire, les meilleures performances sont de l'ordre de 1% de taux d'erreurs grossières (GER). C'est ce GER qui est généralement utilisé comme critère principal de comparaison de performance. Le GER quantifie le nombre de F_0 de référence que l'AEP n'a pas réussi à correctement estimer selon une certaine tolérance. Cette tolérance est différente entre le domaine musical et la parole. En parole, la valeur typique est de 20% d'écart relatif alors qu'en musique elle est de 3%. Les 3% de tolérance du domaine musical s'expliquent facilement par le fait que l'espace des fréquences est discrétisé en notes espacées d'un demi-ton (6%). Ces notes n'existant pas en parole (hormis la voix chantée qui n'est pas traitée dans ce papier), une tolérance de 3% n'est pas nécessairement la plus adaptée. Le choix de 20% peut sembler élevé mais s'explique par le fait que la grande majorité des erreurs d'estimation de F_0 sont des erreurs d'octave ou de sous-octave.

Les évaluations ne donnent pas souvent la performance des AEP dans la détection de F_0 (décision VnV). Or, l'estimation de F_0 est dépendante de cette décision. Cette dépendance est différente selon les AEP. Dans certains cas, la décision VnV est préalable à l'estimation de F_0 (ex. YIN, SWIPE, PAL présentés dans la section 3), dans d'autres cas, la décision VnV est entremélée à l'estimation de F_0 au sein d'un processus de suivi de F_0 (programmation dynamique ou autre) comme dans PRAAT [2] [1] par exemple. L'interaction entre la détection de F_0 et l'estimation de F_0 est un fait connu mais son influence sur le GER est

1. Musical Information Retrieval Evaluation eXchange : `http://www.music-ir.org/mirex/`
2. Fonction classique *To pitch (ac)...*

très souvent sous-estimée. Si la décision VnV n'est pas identique pour les AEP comparés, l'équité de l'évaluation comparative n'est plus garantie et le taux de GER ne suffit plus à la comparaison des AEP.

L'article est organisé de la manière suivante : la section 2 présente les métriques classiques utilisées pour l'évaluation d'AEP. La section 3 détaille les corpus, les AEP et la manière dont les références F_0 ont été calculées. Dans la section 4, l'influence de la décision VnV sur l'estimation de F_0 est présentée. Dans la section 5, ce point crucial est justifié par l'intermédiaire de deux méthodologies d'évaluation, la première ne donnant pas d'informations sur la qualité de la décision VnV et la seconde uniformisant d'emblée cette décision. L'objectif de ce papier est double : d'une part il s'attache à démontrer qu'il est impératif de prendre en compte l'interaction entre décision VnV et estimation de F_0 et d'autre part, il se propose de clarifier les manières de procéder pour réaliser une évaluation comparative équitable.

2. Métriques d'évaluation

Les métriques présentées dans cette section sont proposées dans l'article [9]. Deux processus sont à évaluer : la décision VnV qui renvoie à la détection de F_0 et l'estimation de F_0.

La métrique concernant la décision VnV compare la décision VnV faite par la référence et celle faite par l'AEP. Une trame peut prendre deux états : voisé ou non-voisé. Deux types d'erreurs sont possibles : l'erreur de sous-voisé (faux-rejet) dans laquelle la trame de référence est voisée alors que l'AEP est non-voisé, ou bien l'erreur de sur-voisé (fausse alarme) pour laquelle la trame de référence est non-voisée alors que l'AEP la considère voisée. Le taux d'erreurs de sur-voisé OVR est formalisé dans l'équation 1. Il correspond au nombre de trames en erreur de sur-voisé rapporté au nombre de trames de référence non-voisées.

$$OVR = \Omega[R_U \cap H_V]/\Omega[R_U] \quad (1)$$

$\Omega[...]$ désigne l'opérateur cardinal ensembliste. $\cap$ désigne l'opérateur d'intersection ensembliste. R_U désigne l'ensemble des trames de références non-voisées et H_V désigne l'ensemble des trames données voisées par l'AEP (hypothèses). Le taux d'erreurs de sous-voisé est donné dans l'équation 2. Il correspond au nombre de trames en erreur de sous-voisé rapporté au nombre de trames de référence voisées.

$$UVR = \Omega[R_V \cap H_U]/\Omega[R_V] \quad (2)$$

R_V désigne l'ensemble des trames de référence voisées et H_U désigne l'ensemble des trames données non-voisées par l'AEP. Le couple des taux d'erreurs (OVR, UVR) permet de connaître le point de fonctionnement d'un AEP en termes de décision voisé/non-voisé.

La métrique concernant l'estimation de F_0 classiquement utilisée est le taux d'erreurs grossières noté GER. Il s'agit du nombre de trames dont la valeur F_0 d'hypothèse est distante de plus de 20% en écart relatif absolu de la valeur F_0 de référence. Le GER est calculé sur les trames qui sont déclarées voisées par la référence et par l'AEP. Le GER est donné dans l'équation 3. R_E désigne l'ensemble des trames de référence mal estimées.

$$GER = \Omega[R_E]/\Omega[R_V \cap H_V] \quad (3)$$

3. Matériau

Cette section détaille les matériaux indispensables à toute évaluation comparative d'AEP : les corpus utilisés, la manière dont sont générées les valeurs F_0 de référence et les AEP comparés.

3.1. Corpus de parole

Trois corpus de parole sont utilisés et totalisent environ 50 minutes de parole. Il s'agit des corpus Bagshaw, Keele et Mocha, tous trois choisis car ils contiennent les Electro-Glotto-Grammes (EGG) associés à chaque signal permettant de calculer la vérité terrain F_0 ("*groundtruth*"). Ces trois corpus sont comparables en nature de parole : lue et peu expressive. Bagshaw [3] comporte deux locuteurs, un homme et une femme, prononcant chacun 50 phrases courtes. Le corpus contient environ cinq minutes de parole. Keele [4] comporte dix locuteurs, cinq hommes et cinq femmes, prononcant chacun le même énoncé lu ("*The north wind and the sun...*"). Le corpus contient environ cinq minutes de parole. Mocha [5] comporte deux locuteurs, un homme et une femme, prononçant chacun 460 phrases courtes. Le tout forme un corpus contenant 25 minutes de segments voisés.

3.2. Références F_0

Les valeurs F_0 de référence sont extraites de manière automatique à partir de l'EGG. Pour cela, un algorithme simple utilisant l'autocorrélation est employé. Un post-traitement correctif permet de supprimer les erreurs d'octave ou de sous-octave. Les résultats automatiques obtenus ont été vérifiés manuellement afin de contrôler leur validité. Cette annotation automatique a pour avantage de s'affranchir du travail d'annotation manuel qui peut être fastidieux.

3.3. AEP évalués

Trois AEP sont évalués. Ce choix est justifié par le besoin de paramétrer à volonté la décision VnV ce qui n'est pas toujours le cas selon les AEP. Ils sont également tous purement trame-à-trame (pas de post-traitement). YIN [4] est un algorithme temporel s'appuyant sur la *squared difference function* ou SDF décrite dans l'équation 4. YIN produit des fonctions de pitch dans lesquelles la fréquence du minimum local de plus faible amplitude est considérée comme hypothèse F_0. $x(n)$ est un signal discret de N échantillons. τ varie entre 0 et $N-1$.

$$SDF(\tau) = \sum_{n=0}^{N-1} [x(n) - x(n+\tau)]^2 \quad (4)$$

SWIPE [2] et PAL [7] sont des algorithmes fréquentiels qui calculent le produit scalaire entre un spectre

3. www.cstr.ed.ac.uk/research/projects/fda
4. www.liv.ac.uk/Psychology/hmp/projects/pitch.html
5. www.cstr.ed.ac.uk/research/projects/artic/mocha.html

d'amplitude (noté $|X|$) et une fonction noyau (notée K). Pour SWIPE, cette fonction est une ondelette cosinusoïdale décroissante et pour PAL, il s'agit du peigne alterné [7]. Les fonctions de pitch produites (notée Φ) quantifient une force de périodicité de chaque fréquence dans l'intervalle de recherche. Le maximum local le plus fort est considéré comme hypothèse F_0. L'équation 5 donne la valeur de la fonction de pitch pour une fréquence F_c donnée.

$$\Phi(F_c) = \sum_{n=0}^{N-1} |X(nF_c)|.K(F_c, f) \quad (5)$$

4. Décision voisé/non-voisé et GER

Cette section montre l'impact de la décision VnV sur le taux de GER. En effet, si l'AEP est paramétré de manière à ne considérer que les trames dont le voisement est fort, alors l'estimation de la F_0 sur ces trames est probablement correcte. Par contre, si l'algorithme évalué estime la F_0 sur n'importe quelle trame, alors il considère des trames dont le voisement est plus litigieux et sera davantage sujet aux erreurs. La figure 1 montre ce comportement sur l'algorithme SWIPE. Le comportement est similaire pour les deux autres algorithmes. Ces valeurs sont obtenues sur les trois corpus comptant 25 minutes de parole voisée (cf. section 3). Comme attendu, la probabilité d'erreur augmente fortement lorsque la force de périodicité diminue. Cette probabilité représente le nombre d'erreurs à 20% rapportées au nombre total de valeurs à estimer (correctes+erreurs) pour une force de périodicité donnée. Or, la manière de calculer la force de périodicité diffère d'un algorithme à un autre. Ainsi, le comportement de chacun des algorithmes quant à la décision VnV est différent. Dans ce contexte de théorie de la décision, les courbes ROC ou DET sont classiquement utilisées. Les courbes DET [8] données figure 2 illustrent les différences des trois AEP en termes de détection de F_0 sur le corpus de Mocha. L'objectif n'étant pas ici de donner un classement des AEP, les trois algorithmes sont rendus anonymes et remplacés par A, B et C.

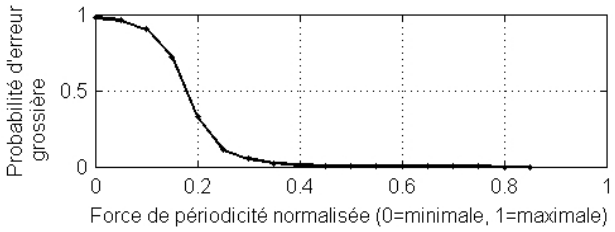

Figure 1: Probabilité d'erreurs grossières en fonction de la force de périodicité.

Il reste enfin à montrer l'influence de la décision VnV sur l'estimation de F_0. Ce point est l'objet de la figure 3. Les courbes présentent le taux GER en fonction du point de fonctionnement de la décision VnV pour les trois AEP. Comme attendu, lorsqu'un AEP favorise le sous-voisement il fait moins d'erreurs grossières (OVR/UVR inférieur à 1). Ainsi, il est possible d'être artificiellement meilleur que d'autres AEP si le point de fonctionnement de l'AEP testé lui est favorable. Mais cette supériorité n'est que la conséquence d'une sélection implicite de trames "bien voisées". Or, c'est justement dans le traitement des trames litigieuses que se manifeste la qualité véritable d'un AEP. Pour éviter ce biais, il est nécessaire soit de donner systématiquement les taux d'OVR et d'UVR, soit de mettre en œuvre une méthodologie d'évaluation qui uniformise les OVR et UVR pour tous les AEP comparés.

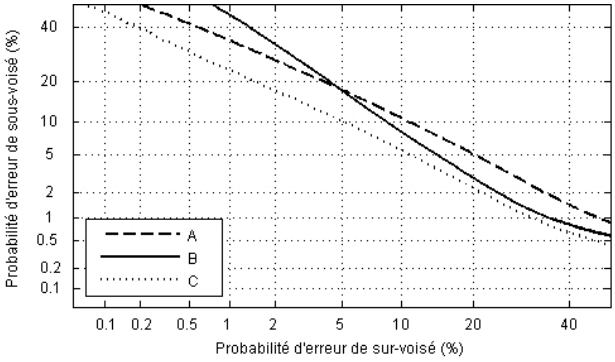

Figure 2: Courbe DET obtenue sur Mocha.

5. Evaluation comparative

Cette section permet de confirmer par l'expérience ce qui a été vu dans la section 4. Pour cela deux méthodologies d'évaluation sont utilisées. La première est nommée méthodologie *standard* et la seconde est nommée méthodologie *sans sous-voisé*. Cette section montre que les résultats bruts de la méthodologie standard sont à prendre avec précaution et qu'il faut toujours mettre en parallèle les résultats d'OVR et UVR avec le GER.

5.1. Méthodologies

Standard (STD) Utilisée dans [2], elle consiste à évaluer les AEP dans les versions préconisées par les auteurs. Le principal avantage est sa simplicité et le fait de pouvoir la mettre en place quel que soit l'AEP (boîtes noires). La décision VnV n'est pas uniforme selon les AEP. Ceci implique que les trames évaluées ne sont pas forcément les mêmes. A la limite, les résultats obtenus par deux AEP différents pourraient être les mêmes alors que les trames évaluées sont toutes différentes. Pour être complète et équitable, cette méthodologie **doit** renseigner le taux d'erreurs de sous-voisé et le taux d'erreurs de sur-voisé.

Sans sous-voisé (SSV) Utilisée dans [3], elle est la plus équitable et consiste à régler la décision VnV des AEP comparés de manière à ce qu'aucune erreur de sous-voisé ne soit faite. Ainsi tous les AEP sont évalués sur les mêmes trames et ces trames sont les trames voisées de la référence. La méthodologie impose un taux d'UVR à 0% et un taux d'OVR à 100%. Cette méthodologie n'est pas toujours applicable car elle nécessite de pouvoir régler la décision VnV, ce qui n'est pas toujours le cas.

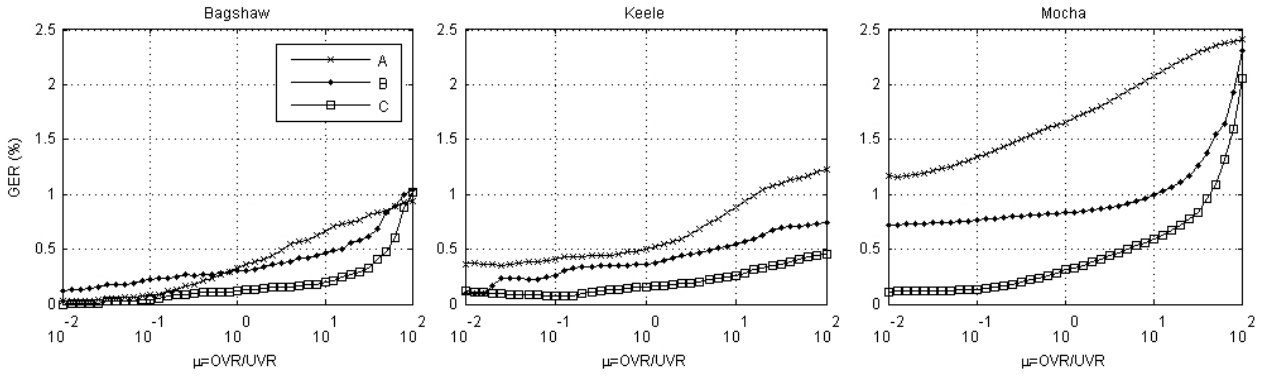

Figure 3: Évolution du GER en fontion du rapport OVR/UVR. L'échelle des abcisses est semi-logarithmique. L'abcisse 10^0 correspond à l'*Equal Error Rate*.

5.2. Résultats

Les résultats obtenus sur les 25 minutes de parole voisée des trois corpus sont récapitulés dans la table 1. Les deux méthodologies sont appliquées. Pour la méthodologie standard, les seuils de décision VnV sont fixés aux valeurs préconisées par les auteurs : 0.2 pour YIN et SWIPE et 1 pour PAL. L'intervalle de recherche de la F_0 est $[60, 500Hz]$.

Table 1: Récapitulatifs des résulats obtenus.

	STD			SSV
	OVR(%)	UVR(%)	GER(%)	GER(%)
A	3.99	21.56	0.38	1.36
B	18.76	3.03	0.45	0.76
C	11.97	2.72	0.19	0.53

Dans la méthodologie standard, les taux d'erreurs grossières sont au-dessous de 0.5%. Le point de fonctionnement ($\mu = OVR/UVR$) de A vaut 0.19, celui de B vaut 6.19 et celui de C vaut 4.40. B est l'algorithme qui donne les moins bonnes performances. Néanmoins, ce résultat doit être considéré avec les taux d'erreur OVR et UVR pour être complet car les μ sont différents. A favorise le sous-voisement contrairement à B et C. Il est donc avantagé dans son taux de GER. C'est effectivement ce qui transparait dans la colonne SSV puisqu'il apparait que le B n'est plus l'algorithme donnant la plus faible performance. B est légèrement moins performant que C alors que A fait environ deux fois plus d'erreurs grossières que les deux autres algorithmes.

6. Conclusions

Cet article montre clairement l'influence de la décision voisé/non-voisé (détection de F_0) sur le taux d'erreurs grossières (estimation de F_0). Il est indispensable de donner les taux d'OVR et d'UVR reflétant la décision VnV des AEP. Si seul le GER est fourni, l'équité de l'évaluation s'en trouve affectée. Ce travail montre que la qualité d'un AEP ne se limite pas à une pure estimation de F_0 mais aussi à la décision VnV. Il y a donc deux axes d'améliorations des AEP : travailler sur l'estimation de F0 et travailler sur la décision VnV. Ce dernier point soulève le problème de vérité terrain en termes de décision voisé/non-voisé et de quantification de la force de voisement.

Références

[1] Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *IFA 1993*, volume 17, pages 97–110, 1993.

[2] Arturo Camacho and John G. Harris. A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, 124(3) :1638–1652, 2008.

[3] Alain de Cheveigne and Hideki Kawahara. Comparative evaluation of f0 estimation algorithms. In *Eurospeech 2001*, volume 4, pages 2451–2454, 2001.

[4] Alain de Cheveigne and Hideki Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4) :1917–1930, 2002.

[5] Nuno Fonseca and Anibal Ferreira. Measuring music transcription results based on a hybrid decay/sustain evaluation. In *ESCOM 2009*, pages 119–124, Finland, 2009.

[6] Wolfgang Hess. *Pitch Determination of Speech Signals : Algorithms and Devices.* Springer-Verlag, Germany, heidelberg edition, 1983.

[7] Jean-Sylvain Liénard, François Signol, and Claude Barras. Speech fundamental frequency estimation using the alternate comb. In *Interspeech 2007*, Antwerpen, Belgium, 2007.

[8] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Eurospeech 1997*, pages 1895–1898, 1997.

[9] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal. A comparative performance study of several pitch detection algorithms. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(5) :399–418, 1976.

[10] Peter Veprek and Michael S. Scordilis. Analysis, enhancement and evaluation of five pitch determination techniques. *Speech Communication*, 37(3-4) :249–270, 2002.