Les temps de traitement des voix de femmes et d'hommes sont-ils équivalents ?

Erwan Pépiot

Les temps de traitement des voix de femmes et d'hommes sont-ils équivalents ?

Erwan Pépiot

Groupe LAPS – EA1569
Université Paris 8 - 2 rue de la liberté 93200 Saint-Denis
erwan.pepiot@free.fr

RESUME

Cette étude a pour objet les temps de traitement des voix de femmes et d'hommes. Plusieurs auteurs ont mis en évidence la difficulté accrue de l'identification des voyelles lorsque ces dernières sont produites avec un F0 élevé (Ryalls & Lieberman, 1982). Cela a-t-il des conséquences sur le traitement des mots ? Les voix de femmes sont-elles traitées plus lentement que les voix d'hommes ? Une expérience de détection de mots a été réalisée, afin de tester le temps de réponse des participants en fonction du genre du locuteur ayant produit le mot-cible. Les résultats suggèrent que les voix d'hommes et de femmes sont traitées par l'auditeur à vitesse équivalente, mais néanmoins comme deux entités différentes.

ABSTRACT

Are female and male voices processed equally fast?

This study deals with processing time of female and male speech. Several authors showed that vowel identification was more difficult on voices with a high F0 (Ryalls & Lieberman, 1982). Does this have consequences on word processing? Are female voices processed more slowly than male ones? A word spotting experiment was conducted in order to test the participants' response time, depending on whether the target word is produced by a male or a female voice. Results suggest that these two types of voice are processed equally fast, even though they seem processed as two different entities.

MOTS-CLES : voix de femmes, voix d'hommes, temps de traitement, détection de mots.
KEYWORDS : female voices, male voices, processing time, word spotting.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 153–160,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Les voix de femmes ont souvent été négligées par les phonéticiens, en particulier dans les études impliquant des relevés formantiques. Cela s'explique en partie par le fait que leurs formants vocaliques sont généralement plus durs à localiser que ceux de leurs homologues masculins sur les spectrogrammes et les spectres. En effet, les locuteurs féminins disposant d'un F0 moyen globalement plus élevé que celui des hommes, leurs voix présentent moins d'harmoniques, entraînant ainsi des formants moins repérables. Qu'en est-il alors du traitement de ces deux types de voix par l'auditeur ?

(Sokhi & al., 2005) et (Lattner & al., 2005) ont montré, grâce à l'utilisation de l'IRMf, que l'écoute de voix d'hommes et de femmes n'activait pas de la même manière certaines zones du cerveau de l'auditeur. Ces résultats semblent valider l'hypothèse d'un traitement différencié des voix de femmes et d'hommes par le cerveau.

Concernant la difficulté de traitement, une première étude importante a été réalisée par (Ryalls & Lieberman, 1982). Des voyelles isolées synthétisées ont été présentées à des auditeurs ayant pour tâche de les identifier. Le F0 des voyelles était soit de 100 Hz, de 135 Hz, ou de 250 Hz. Dans tous les cas, les voyelles avec un F0 à 100 Hz et à 135 Hz ont été significativement mieux identifiées par les auditeurs que celles avec un F0 à 250 Hz. Selon ces auteurs, une voix présentant beaucoup d'harmoniques (i.e. un F0 bas) facilite la localisation des formants par l'auditeur et donc l'identification des voyelles. Une étude similaire menée par (Diehl et al., 1996) présente les mêmes conclusions.

Compte tenu de ces constatations, on pourrait penser que le temps de traitement des voix est proportionnel au F0 de la voix traitée, donc supérieur pour les voix de femmes. Dans une étude réalisée sur des anglophones américains, (Strand, 2000) a diffusé des mots isolés produits par un homme ou une femme à des participants ayant pour tâche de répéter le mot le plus rapidement possible. Les temps de réponse ont été mesurés et comparés dans quatre conditions : voix d'homme stéréotypique, voix de femme stéréotypique, voix d'homme non-stéréotypique (i.e. ambiguë) et voix de femme non-stéréotypique. Aucune différence significative n'est apparue entre la voix d'homme et la voix de femme stéréotypiques. Les voix ambiguës (femme et homme) ont sans surprise entraîné un temps de réaction significativement plus long.

Cette étude suggère donc que les voix de femmes ne seraient pas plus longues à traiter par le cerveau que celles des hommes. Cependant, seules une voix d'homme et une voix de femme stéréotypiques ont été utilisées, ce qui ne permet pas de tirer des conclusions générales. De plus, plutôt que le paradigme de répétition de mots, celui de la détection de mots est potentiellement plus révélateur car il implique uniquement un travail de perception. Une expérience de ce type à donc été menée, pour tenter de vérifier les deux hypothèses ci-dessous.

Hypothèse 1 : toutes choses égales par ailleurs, les voix de femmes et les voix d'hommes sont traitées à vitesse équivalente.

Hypothèse 2 : les voix de femmes et les voix d'hommes sont considérées comme deux entités distinctes par le cerveau.

2 Méthode

2.1 Matériau linguistique et enregistrements

La détection de mots est un paradigme expérimental qui se caractérise par la diffusion de *séries de mots* de longueurs variables et se terminant par un *mot-cible*, préalablement communiqué au participant (Marslen-Wilson & Tyler 1980). La tâche de ce dernier est d'appuyer sur un bouton dès qu'il perçoit ce mot-cible.

Le choix des mots a été réalisé sur la base de plusieurs critères : une longueur équivalente (dissyllabiques), une fréquence d'occurrence élevée (figurant dans les 1500 mots les plus fréquents de la langue française[1]), un contenu émotionnel le plus neutre possible. Au total, 61 mots différents ont été sélectionnés, soit 1 mot-cible et 60 autres mots. La cible choisie est le mot *étage*. Ce dernier a été retenu en raison de sa voyelle initiale [e], présentant d'importantes différences formantiques entre hommes et femmes, et donc susceptible de maximiser les effets recherchés dans cette expérience.

Pour ces enregistrements, j'ai fait appel à 8 locuteurs francophones : 4 femmes et 4 hommes, âgés de 20 à 34 ans. Tous sont locuteurs du français dit *parisien*, non-fumeurs et ne présentant pas de trouble de la parole. Les enregistrements ont été effectués en chambre sourde, à l'aide d'un enregistreur numérique. Chaque locuteur a été enregistré sur l'ensemble des 61 mots. Afin d'homogénéiser les paramètres prosodiques, chaque mot a été placé dans le contexte suivant : « *Il a dit* MOT *deux fois* ». Ces mots ont par la suite été extraits de leur contexte.

2.2 Participants

Au total, 25 auditeurs (8 hommes et 17 femmes) ont pris part à cette expérience. Ces participants sont tous des francophones natifs ne présentant pas de troubles du langage et âgés de 18 à 65 ans. La moyenne d'âge est de 27,6 ans : 36,1 ans pour les hommes, 23,6 ans pour les femmes.

2.3 Procédure expérimentale

Une expérience de détection de mots nécessite la maîtrise de plusieurs variables inhérentes à ce paradigme, et la neutralisation de divers biais. Quatre conditions expérimentales doivent être utilisées pour tester les hypothèses :

- **Condition A** (*homogène*) : contexte *voix d'hommes* avant mot-cible expérimental *voix d'homme*.
- **Condition B** (*homogène*) : contexte *voix de femmes* avant mot-cible expérimental *voix de femme*.
- **Condition C** (*non-homogène*) : contexte *voix de femmes* avant mot-cible expérimental *voix d'homme*.
- **Condition D** (*non-homogène*) : contexte *voix d'hommes* avant mot-cible expérimental *voix de femme*.

[1] Sur la base de données mises à disposition par le Ministère de l'Education Nationale, de la Jeunesse et de la Vie Associative : http://eduscol.education.fr/cid47916/liste-des-mots-classee-par-frequence-decroissante.html.

Afin de maximiser son effet, le *contexte* s'étend non seulement sur les 4 mots non-cibles de la série expérimentale, mais également sur la série précédente, que je nommerai *pré-expérimentale* et longue de 3 à 4 items, mot-cible inclus, *soit un total de 7 à 8 mots précédant directement le mot-cible expérimental*. Des séries de distracteurs, contenant chacune un mot-cible mais dont les temps de réponse ne seront pas pris en compte, ont également été utilisées et un schéma de base est ainsi répété : deux séries de distracteurs, une série pré-expérimentale, une série expérimentale. La longueur des séries de distracteurs varie de 2 à 7 items, mot-cible inclus.

L'expérience se divise en quatre *blocs* de 16 séries de mots, comportant à chaque fois les quatre conditions expérimentales dans un ordre différent :

- **Bloc 1** : (2 séries de distracteurs), Cond. A, (2 séries de distracteurs), Cond. B, (2 séries de distracteurs), Cond. C, (2 séries de distracteurs), Cond. D.
- **Bloc 2** : (2 séries de distracteurs), Cond. B, (2 séries de distracteurs), Cond. A, (2 séries de distracteurs), Cond. D, (2 séries de distracteurs), Cond. C.
- **Bloc 3** : (2 séries de distracteurs), Cond. D, (2 séries de distracteurs), Cond. C, (2 séries de distracteurs), Cond. B, (2 séries de distracteurs), Cond. A.
- **Bloc 4** : (2 séries de distracteurs), Cond. C, (2 séries de distracteurs), Cond. D, (2 séries de distracteurs), Cond. A, (2 séries de distracteurs), Cond. B.

Chaque *condition*, qui contient deux séries de mots (*pré-expérimentale* et *expérimentale*), est donc testée quatre fois dans l'expérience, en occupant chacune des positions possibles (1, 2, 3 ou 4) à l'intérieur des blocs.

Un même mot-cible, le mot *étage*, a été utilisé pour toute l'expérience, afin de limiter les divers biais qu'aurait pu induire l'utilisation de mots-cibles variés. Par conséquent, un autre élément a dû être pris en compte : à force de détecter un même mot-cible, il est possible que les auditeurs améliorent globalement leur temps de réponse au fur et à mesure qu'ils avancent dans l'expérience. Pour compenser cet éventuel biais, une moitié d'auditeurs s'est vu diffuser les blocs dans l'ordre 1, 2, 3, 4 et l'autre moitié dans l'ordre 3, 4, 1, 2. De plus, une vérification statistique sera réalisée *a posteriori*.

Tous les mots non-cibles apparaissent une fois par bloc, toujours dans un ordre différent. Quant aux séries d'entraînement, diffusées en début d'expérience, 7 mots spécifiques ont été utilisés, chacun apparaissant 3 fois. La répartition des voix pour les différents mots s'est faite selon plusieurs règles, établies en vue de limiter de façon optimale les différents biais possibles. Ainsi, sur l'ensemble de l'expérience, chaque voix apparaît deux fois en position de mot-cible expérimental et aucune voix n'apparaît plus de deux fois dans une même série (y compris expérimentale), ni sur deux mots consécutifs.

L'expérience à été réalisée à l'aide d'un ordinateur portable, du logiciel *Perceval 3.0.5.0* et d'un boîtier externe : l'utilisation de ce type de périphérique a l'avantage de permettre une mesure très précise des temps de réponse. Une fois installé devant l'écran d'ordinateur et équipé d'un casque audio, le participant était invité, par consigne écrite affichée à l'écran, à *appuyer sur le bouton bleu du boîtier le plus rapidement possible dès qu'il entendrait le mot* étage.

Dans un premier temps, six séries de mots d'entraînement étaient diffusées, suivies des quatre blocs constitutifs de l'expérience. Durant toute la durée du test, aucun stimulus

visuel n'était diffusé à l'écran. Les stimuli audio (i.e. les mots) ont été présentés avec un intervalle inter-stimulus de 600 ms. Le choix de cet intervalle relativement court a été effectué dans le but de maintenir éveillée l'attention du sujet tout au long de l'expérience tout en limitant la durée totale de celle-ci.

3 Analyse des données

Les temps de réponse, c'est-à-dire le délai entre le début du mot-cible et l'appui sur le bouton par le participant, étaient automatiquement inscrits par *Perceval* dans un fichier texte. Au total, 64 temps de réponse par participant ont été collectés, correspondant à tous les mots-cibles de l'expérience (hors séries d'entraînement), qu'ils apparaissent dans des séries de distracteurs, pré-expérimentales ou expérimentales. *Seuls les temps de réponse correspondant aux séries expérimentales ont été effectivement retenus*. Parmi ces derniers, aucune mesure pouvant être considérée comme « aberrante » n'a été observée : tous ces temps de réponse ont donc été conservés et pris en compte pour les résultats. Seize mesures ont ainsi été relevées par participant (4 par condition expérimentale). Pour l'ensemble des 25 participants, cela correspond à un total de 400 mesures, soit 100 temps de réponse pour chacune des 4 conditions expérimentales.

4 Résultats

Les temps de réponse moyens obtenus pour la reconnaissance du mot-cible dans les quatre conditions expérimentales, pour les 25 auditeurs, sont les suivants :

- **Condition A** (*homogène, mot-cible voix d'homme*) : 502 ms.
- **Condition B** (*homogène, mot-cible voix de femme*) : 495 ms.
- **Condition C** (*non-homogène, mot-cible voix d'homme*) : 478 ms.
- **Condition D** (*non-homogène, mot-cible voix de femme*) : 474 ms.

Les temps de réponse moyens sont relativement proches entre les conditions A et B d'une part, et entre les conditions B et C d'autre part, c'est-à-dire entre les mots-cibles produits par des hommes et ceux produits par des femmes, en contexte équivalent. Une différence assez importante apparaît en revanche entre les conditions A et C, ainsi que B et D (temps de réponse plus courts dans les conditions C et D), laissant supposer un possible effet du contexte (homogène ou non-homogène avec le mot-cible) sur les temps de traitement des mots-cibles.

Afin de vérifier diverses interactions possibles entre les facteurs et d'établir si les différences constatées sont significatives, j'ai procédé à une analyse statistique des résultats à l'aide du logiciel *StatView 5.0*.

Dans un premier temps, j'ai souhaité m'assurer que le genre des auditeurs n'avait pas eu d'influence sur les temps de réponse obtenus en fonction des différentes conditions expérimentales. Le résultat de l'ANOVA est clair : il n'existe aucune interaction entre les facteurs « genre des auditeurs » et « condition expérimentale » ($F(3,392) = 0,299$; $p > 0,80$). Cela suggère que *les différences relatives de temps de réponse entre les quatre conditions expérimentales (A, B, C, D) n'ont pas varié en fonction du genre des auditeurs.* L'analyse des temps de réponse en fonction des conditions expérimentales pourra donc être effectuée sur l'ensemble des auditeurs, indépendamment de leur genre.

Un autre biais potentiel existe : la longueur du mot-cible *étage*, qui varie sensiblement en fonction du locuteur l'ayant produit. J'ai donc effectué un test de Pearson sur le temps de réponse moyen des auditeurs et la durée des stimuli. Il en est ressorti une très faible corrélation : r(8) = 0,206. Cette dernière est très largement non significative, avec $z = 0{,}467$ et $p > 0{,}60$. *La longueur du mot-cible ne semble donc pas avoir joué sur les temps de réponse des auditeurs.*

Comme cela a été mentionné précédemment, l'utilisation d'un même mot-cible tout au long de l'expérience aurait pu entraîner une diminution progressive du temps de réponse des sujets. Un test de Spearman a été conduit sur les temps de réponse des sujets et le moment de diffusion de chaque mot-cible expérimental. On observe une absence totale de corrélation entre ces deux variables ($rhô = 0{,}001$; $p > 0{,}95$) : *la répétition du mot-cible* étage *ne semble donc pas avoir entraîné de diminution des temps de réponse des auditeurs.*

Les possibles biais ayant été écartés, j'ai ensuite testé l'effet du facteur « condition expérimentale », en effectuant une ANOVA à deux facteurs : « condition expérimentale » et « sujet » (ce deuxième facteur a été inclus afin d'obtenir une variance plus juste), sur les temps de réponse des auditeurs. Le graphique correspondant à cette analyse est visible ci-dessous (Figure 1).

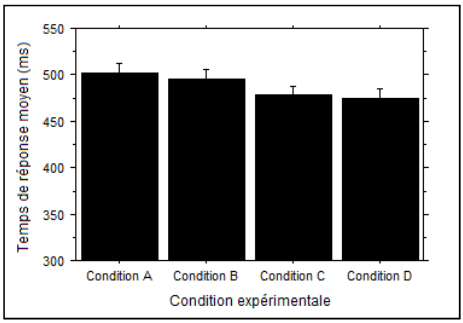

FIGURE 1 – Temps de réponse moyens (ms) des auditeurs en fonction de la condition expérimentale, avec les barres d'erreur correspondantes (± 1 erreur-type).

Le résultat obtenu montre l'existence d'un effet global significatif du facteur « condition expérimentale » ($F(3,300) = 4{,}597$; $p < 0{,}01$). *Globalement, les temps de réponse des auditeurs ont donc varié significativement en fonction de la condition expérimentale.*

De manière plus précise, le test PLSD de Fisher révèle que la différence est significative entre les conditions A et C ($p < 0{,}01$), ainsi qu'entre les conditions B et D ($p < 0{,}02$). Il existe donc un effet du contexte : *les temps de réponse pour les voix de femmes, comme pour les voix d'hommes, ont été significativement plus faibles dans les conditions non-homogènes (C et D), où les mots précédant le mot-cible sont produits par des voix du genre opposé, que dans les conditions homogènes (A et B)*. En revanche, les différences entre les conditions A et B ($p > 0{,}40$), et C et D ($p > 0{,}60$) ne sont pas significatives. *En contexte équivalent, les mots-cibles produits par des femmes et ceux produits par des hommes ont donc entraîné des temps de réponse similaires.*

En plus de cette comparaison en contexte équivalent, j'ai souhaité vérifier si, globalement, les temps de réponse moyens obtenus pour les voix de femmes (conditions B et D) et pour les voix d'hommes (conditions A et C) ne présentaient pas de différence significative. Pour cela, j'ai regroupé les temps de réponse des conditions B et D (voix de femmes), et ceux des conditions A et C (voix d'hommes), et effectué une ANOVA à deux facteurs, « type de voix produisant le mot-cible » et « sujet », sur le temps de réponse moyen des auditeurs. Le graphique représentant les temps de réponse moyens en fonction du type de voix produisant le mot-cible est visible ci-après (Figure 2).

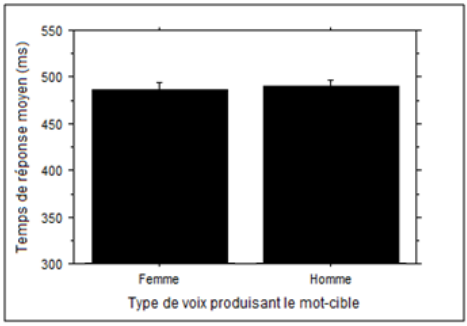

FIGURE 2 – Temps de réponse moyens (ms) des auditeurs, en fonction du type de voix (femme ou homme) produisant le mot-cible, avec les barres d'erreur (± 1 erreur-type).

Il n'existe aucun effet significatif du facteur « type de voix produisant le mot-cible » sur les temps de réponse moyens ($F(1,350) = 0,852$; $p > 0,30$). *Globalement, il n'y a donc pas de différence significative entre les temps de réponse des auditeurs pour les mots-cibles produits par des femmes et pour ceux produits par des hommes.*

5 Discussion - Conclusion

Cette expérience de détection de mots a permis d'obtenir certains résultats intéressants. Tout d'abord, l'hypothèse 1 a été confirmée : aussi bien en contexte équivalent que tous contextes confondus, les temps de traitement des mots produits par des hommes et de ceux produits par des femmes ne présentent aucune différence significative. Ainsi, *les voix de femmes et les voix d'hommes semblent être traitées à la même vitesse par les auditeurs sur les mots isolés*, et ceci *indépendamment du genre de l'auditeur*.

Ce résultat est à mettre en perspective avec des recherches antérieures. (Ryalls & Lieberman, 1982) et (Diehl & al., 1996) avaient mis en évidence le lien entre F0 et difficulté d'identification des voyelles. Cela pouvait suggérer que les voix de femmes sont plus difficiles à traiter par les auditeurs. Mais ces expériences ont mesuré le pourcentage d'erreur d'identification et non pas le temps de réponse. D'autre part, l'unité linguistique utilisée (voyelle isolée) peut sembler quelque peu artificielle : en dehors de conditions expérimentales, les auditeurs ont rarement à identifier une unité si petite hors contexte. On peut donc penser que les auditeurs, qui sont quotidiennement exposés à des voix à F0 élevé ainsi qu'à des voix à F0 bas, ont pu développer des capacités de traitement

similaires pour ces différents types de voix, pour un input d'une taille au moins équivalente à celle d'un mot. Ainsi, même si les voyelles produites avec un F0 élevé sont plus difficiles à identifier, les auditeurs ont la possibilité de compenser avec les consonnes, dont on sait qu'elles sont particulièrement décisives pour l'accès au lexique (Owren & Cardillo, 2006).

(Strand, 2000) a quant à elle utilisé des mots isolés et mesuré les temps de réponse en fonction du type de voix, comme dans la présente étude. Néanmoins, le paradigme utilisé était une tâche de répétition de mots, ce qui implique non seulement une tâche de perception mais également un travail de production. Malgré ces divergences méthodologiques, les résultats obtenus dans l'expérience de Strand sont conformes à ceux obtenus ici : aucune différence significative de temps de réponse n'a été observée entre voix de femmes et voix d'hommes. Notons que dans cette précédente étude, seule une voix d'homme et une voix de femme dites « stéréotypiques » avaient été utilisées : il était donc nécessaire de confirmer ces tendances avec un plus grand nombre de voix.

La deuxième observation importante concerne les différences obtenues entre les conditions A et C d'une part et B et D d'autre part : les conditions *non-homogènes* (C et D) ont entraîné des temps de traitement inférieurs à celles dites *homogènes* (A et B). L'écoute d'un grand nombre de stimuli de type *voix d'homme* avant un mot-cible de type *voix de femmes* (ou *vice versa*) a fait baisser le temps de réponse des auditeurs. Cela s'explique probablement par un regain d'attention du sujet dû à un changement de paradigme, et semble donc aller dans le sens de l'hypothèse 2 selon laquelle les voix de femmes et d'hommes sont considérées comme deux entités distinctes par le cerveau. Ces résultats paraissent confirmer ceux obtenus par (Sokhi & al., 2005) et (Lattner & al., 2005), montrant que ces deux types de voix activent de manière différente certaines zones du cerveau des auditeurs.

Références

DIEHL, R. L. et al. (1996). On explaining certain male-female differences in the phonetic realization of vowel categories. *Journal of Phonetics*, 24, pages 187–208.

LATTNER, S. & al. (2005). Voice perception: Sex, pitch, and the right hemisphere. *Human Brain Mapping*, 24, pages 11–20.

MARSLEN-WILSON, W. & TYLER, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, 8, pages 1–71.

OWREN, M. et CARDILLO, G. (2006). The relative roles of vowels and consonants in discriminating talker identity versus word meaning. *Journal of the Acoustical Society of America*, 119, pages 1727–1739.

RYALLS, J. H. et LIEBERMAN, P (1982). Fundamental frequency and vowel perception. *Journal of the Acoustical Society of America*, 72, pages 1631–1634.

SOKHI, D. S et al. (2005). Male and female voices activate distinct regions in the male brain. *NeuroImage*, 27, pages 572–578.

STRAND, E. (2000) *Gender stereotype effects in speech processing*. PhD Thesis. The Ohio State University.