Expressivité et synthèse vocale. Isotopies expressives, cohérence discursive et structures prosodiques

Ioana Suciu[1&2], Ioannis Kanellos[2], Thierry Moudenc[1]

[1] TECH\SSTP\VMI, France Télécom R&D, Lannion
[2] Département Informatique, ENST Bretagne, Brest
<{ioana.suciu, ioannis.kanellos}@enst-bretagne.fr,
thierry.moudenc@orange-ftgroup.com>

Résumé

Nous traitons dans cet article des rapports de l'expressivité orale au discours, spécifiquement dans la perspective de la synthèse vocale. Après une rapide discussion sur la notion d'expressivité en tant que dimension sémiotique complémentaire du discours et sur les deux linguistiques qu'elle convoque (de la phrase et du texte), nous visitons ses rapports au texte, en cherchant son rôle dans la cohérence sémantique discursive. Nous abordons ensuite la prosodie comme espace d'observables où l'on peut traiter l'expressivité orale de façon formelle. Nous esquissons enfin l'objectif applicatif de notre analyse ainsi que la méthodologie que nous suivons pour rendre la synthèse vocale expressive.

1. Introduction

Verrou technologique important, l'expressivité de la parole convoque aujourd'hui des problématiques théoriques ardues qui réactivent les passions sur la textualité et exigent des formes d'analyse jusqu'alors écartées de la norme des recherches linguistiques. En effet, le développement des théories linguistiques a longtemps suivi les détours d'une réflexion fondée sur l'écrit et limitée essentiellement à la phrase. La voix, en tout rebelle, débordait la compétence linguistique et ne semblait guère se plier aux exigences d'un objet d'étude fiable. Certes, le postulat de la primauté de la parole (sur la langue) ne s'affaiblissait pas ; il se voyait toutefois marginalisé. Par un retour de l'histoire cependant, préparé probablement par les impératifs d'un marché de plus en plus intéressé par le traitement de la langue, l'oral semble regagner sa légitimité. Dans ce mouvement de retour, il amène aussi de nouveaux éclairages aux problèmes traditionnels et ouvre les études linguistiques à des méthodologies inédites, en se rangeant aux côtés d'une linguistique du texte et de l'interprétation et en s'ouvrant à des pratiques fondées sur des corpus.

L'expressivité de la parole en est un cas probablement emblématique. Constituant de l'art verbal, elle ne concerne plus la poétique et la

rhétorique de l'écrit mais réaffirme la nécessité de considérer les objets linguistiques à leur correct niveau d'analyse : celui du texte, mieux, du discours, auquel elle impose désormais de reconnaître une dimension de signifiance complémentaire, portée précisément par l'oral. Et c'est justice ! La primauté de la parole a longtemps été dissociée de la primauté expressive. En apprenant une langue, on apprend, outre son lexique, ses spécificités morpho-syntaxiques, ses figures de discours etc., un ensemble de « schémas » expressifs d'ordre oral, omniprésents à son utilisation, et dont le siège naturel est le texte. On sait aujourd'hui que, avant même les mots et les phrases, c'est le registre expressif oral qui s'impose comme entrée dans la langue, devenant cadre de la compétence linguistique.

Mais que doit-on comprendre par « expressivité orale » ? Bien sûr, ce qui fait d'un texte un discours, un discours porté par une voix. Mais aussi une classe de problématiques de l'émergence du sens. Tout d'abord, l'expressivité reformule le problème sempiternel de la constitution du global. Certes, toute unité d'analyse linguistique (du phonème au discours) peut contenir des éléments expressifs et prétendre, en dernier ressort, à quelque problématique d'expressivité. Mais comment obtient-on l'expressivité des formes globales, notamment celle d'un discours entier ? À partir de quels éléments et par quels moyens ? Puis, s'il est vrai qu'elle se centre sur le pathos, l'expressivité s'affirme généralement dans les conditions du logos, auquel elle est pertinemment consubstantielle. Volontairement découpée du texte, elle ne concerne plus directement la communication linguistique mais d'autres formes d'échange social, qui se posent comme un langage second : elle constitue un espace sémiotique autonome, affranchi de l'écrit.

En cherchant à mettre sur pied des briques technologiques incorporant de l'expressivité, la synthèse vocale ne peut longtemps ignorer ces enjeux, somme tout sémiotiques. Du coup, elle vient s'interroger à côté du linguiste théoricien autant sur les rapports de l'expressivité au discours que sur ce qui la rend observable, et plus avant « tractable » dans les modalités du calcul : la prosodie.

2. Expressivité et discours : isotopies expressives et cohérence

Continu et complexe, l'espace de la sémiose discursive demeure avant tout un espace culturellement circonscrit. Tant en production qu'en réception, l'oralité définit des frontières discursives et expressives qui démarquent l'acceptable de l'inacceptable, le partageable du singulier, le porteur de sens du déconcertant voire de l'insensé. La compétence linguistique pénètre même à l'intérieur de ces départages en cher-

chant des formes de stabilité expressive qui déclenchent des régimes d'interprétation adéquats.

Tout d'abord, en codifiant des normes discursives socialement partagées, l'expressivité permet de situer le discours dans un genre textuel et dans une pratique discursive, de reconstituer ses circonstances de production, ainsi que d'identifier les particularités permanentes (la façon de parler) ou circonstancielles (la maladie, la fatigue, l'ivresse…) de l'idiolecte qui le réalise (Abitbol 2005).

Puis, en jouant le rôle d'un catalyseur interprétatif, elle contribue à la mise en place d'un univers d'attentes et de contraintes pour réguler les possibilités de l'interprétation. Elle devient ainsi une sorte de « guide de lecture » pour l'auditeur, en éclairant ses chemins et en contraignant ses errances en matière d'interprétation. Elle propose aussi des procédures de déchiffrement sémiotique, voire des indices de changement de régime interprétatif, capables d'amender le contenu du texte ou son intention. Elle arrive ainsi soit à faciliter soit à rendre plus difficile l'interprétation du discours, tantôt en le rectifiant, tantôt en le falsifiant. Le texte, certes, pose ; mais c'est l'expressivité qui dispose.

Disposer d'un ensemble de formes expressives initiales est certes essentiel pour situer un discours oral. Il permet sans doute de canaliser l'interprétation, de mettre en évidence certains éléments saillants ou de pertinence… Mais seul, il ne suffit pas. À sa suite, c'est la nature de tels éléments, leur récurrence, leur agencement et leur logique d'organisation au sein du discours oral qui parachèvent sa lecture, par la constitution, précisément, de micro-, méso- et macro-*isotopies expressives* qui se tissent graduellement dans une totalité expressive cohérente (Rastier 1987).

En effet, superposant au texte un espace de signifiances discursives orales, l'expressivité y introduit des dimensions parfois complémentaires, parfois concurrentes, mais jamais indifférentes. Affranchies sémiotiquement, ces signifiances peuvent se trouver, localement ou globalement, tantôt dans un rapport de conformité, de spécification ou d'affinement avec le contenu textuel, tantôt en rapport d'alternative ou d'opposition, voire de contradiction, occasionnellement même de rupture. La négociation des ces deux espaces sémiotiques – du texte et de l'expressivité orale – est une question capitale pour toute démarche herméneutique sur la parole lue. Son enjeu est celui de la cohérence.

On peut soutenir qu'un discours cohérent a au moins une identité sémiotique expressive qui dépend du contexte de la locution (genre textuel, situation discursive, profil du locuteur), et que toute identité

expressive est nécessairement une identité globalement cohérente. Si l'unicité de l'identité expressive d'un discours n'est pas exigée, elle reste pourtant amplement recherchée dans la mesure où les enchaînements des éléments expressifs tendent à suivre la direction désignée préalablement par des présomptions d'isotopie expressive.

Un discours est *expressivement cohérent* lorsque « chaque élément (expressif) saillant exerce une fonction particulière et se lie aux autres pour former un même tout » (Désesquelles 1999). Associés à des éléments issus de divers paliers d'analyse linguistique, les éléments expressifs oraux ne s'enchaînent pas de façon arbitraire mais en vertu du rôle que chacun joue. Ils n'ont « ni le même poids, ni le même relief », ils « constituent des hiérarchies » et « instaurent des ordres » (Caelen-Haumont). L'effet sémantique en est une vue d'ensemble sur le discours qui permet de lui retrouver un sens en même temps « dans et entre » les détails (Rastier 2001). Cette vue d'ensemble justifie précisément la formation d'une progression expressive unifiée et évite que l'ensemble se délite en de multiples expressivités isolées. La réussite d'un discours exige donc d'allier *cohérence* et *expressivité*.

Les dernières réflexions motivent l'introduction de la notion de *forme discursive expressive (fde)* : une structure complexe, ancrée dans la matière textuelle qui porte ses déterminations génériques, situationnelles et idiolectales. Culturellement stable, reconnaissable, interprétable et partageable, une *fde* informe sur le rôle d'un ensemble d'éléments (syntaxiques, sémantiques, morphologiques, rhétoriques…) qui participent à l'identité sémantique du discours (Kanellos et al. 2007).

Devenant *interprétants* d'ordre supérieur, les *fde* assurent aussi le déroulement expressif du discours de deux manières : en stabilisant des isotopies et en tissant des cohérences locales. Pour devenir, *in fine*, des éléments de perception, de compréhension, d'analyse et de reproduction des jeux de périodicités, de structures et de mouvements discursifs sur lesquels se fondent les *rythmes expressifs* (Sauvanet 2000). Autrement dit, d'éléments de réception, ils peuvent muter vers des pratiques de production sémantique.

3. Expressivité et prosodie : niveaux d'analyse, isotopies expressives et vecteurs prosodiques

L'expressivité orale est traditionnellement rationalisée par la *prosodie*, qui est censée traduire l'ensemble des mouvements expressifs, locaux ou globaux, par le jeu de paramètres relevant de trois structures constituantes fondamentales de l'oralité : la mélodie, le tempo et l'énergie. Ces paramètres sont rapportés sur des unités de différents niveaux d'analyse, et dont le choix pour l'étude des phénomènes expressifs est crucial. En effet, discrétisant le continu expressif, les niveaux

d'analyse réduisent la complexité du dire et permettent d'en éclaircir certaines dimensions pour pouvoir les exploiter et les approfondir. Du phonème à l'intertexte, le choix est large et toute tentative d'exhaustivité est condamnée à l'impuissance devant le complexe. Outre le texte, qui reste l'horizon ultime de traitement, les niveaux du *groupe phrastique (gph)*, du *syntagme (syn)* et de la *syllabe (syl)* semblent un bon compromis entre exigence théorique et pragmatisme applicatif. Il y en a des raisons simples.

Alors que le niveau de la phrase ne suffit pas et que le passage directement au palier textuel devient fortuit pour une analyse exploitable, le groupe phrastique semble assurer honnêtement une position intermédiaire. Pouvant s'identifier, selon le cas, avec la première ou avec le second, il est défini comme un groupe de phrases contiguës au sein d'un discours. Qualitativement, il est texte ; quantitativement, une suite de phrases. On remarque ensuite que, à l'intérieur d'un *gph*, la délimitation des syntagmes (tant syntaxiques que rythmiques) joue un rôle capital pour la fluidité et l'intelligibilité de la parole, dans la mesure où « la rupture du syntagme établit la discontinuité sur le dire » (Lafont 1994). Il serait difficile de les omettre. Enfin, on ne saurait omettre non plus la syllabe, unité constitutive des syntagmes et interface naturelle entre le niveau segmental et supra-segmental.

Si les isotopies expressives peuvent se traduire par des déploiements prosodiques d'ordre mélodique, temporel et énergétique, le travail de mise en correspondance entre expressivité et prosodie est loin d'être évident, car il revient à consentir une concordance entre deux espaces de nature différente : le qualitatif expressif et le quantitatif prosodique. Assurément, il existe des cas où les éléments d'ancrage expressif sont soulignés par des saillances prosodiques facilement repérables et où la mise en avant des unités expressives se fait par une différenciation d'intonation, d'énergie, de durée syllabique, de tempo des syntagmes etc. (p.e. l'emphase). Mais leur correspondance prosodique générale est difficilement discernable dans l'amas des jeux de paramètres. En réalité, les interprétants prosodiques changent de portée, s'organisent dans des structures plus amples ou se cachent derrière des unités linguistiques couramment écartées des analyses prosodiques habituelles. Moins saillants, leur identification dépend considérablement de la rationalisation de la finesse de l'écoute.

Dans la danse des sons et des silences donc, des schémas prosodiques s'instaurent et se développent pour faire émerger graduellement une cohérence globale. Les « mouvements expressifs » qu'ils rendent perceptibles peuvent se décrire au moyen de paramètres prosodiques

comme des vecteurs qui se constituent sur trois structures prosodiques imbriquées, correspondant aux trois niveaux d'analyse élus (*gph*, *syn*, *syl*) (Suciu et al. 2006). Ainsi, par des projections convenables des faits expressifs manifestés dans une parole attestée sur ces trois niveaux, les vecteurs prosodiques permettent de capter et de formaliser une information complémentaire nécessaire pour observer, analyser et manipuler des formes expressives du discours.

Cette entreprise de formalité transforme plus avant la représentation prosodique en un « atelier de création » d'expressivité. En effet, de nouvelles formes expressives peuvent être engendrées à tout moment, en suivant la trame des déformations que peuvent subir les vecteurs prosodiques : compressions et étirements, inversions et transgressions, translations et mises en échelle… peuvent opérer sur les composants de vecteurs prosodiques à des niveaux différents pour offrir à la parole de nouvelles physionomies expressives. Cependant, la difficulté de la génération prosodique ne réside pas dans le processus de génération de ces formes, mais plutôt dans leur validation. Les opérateurs formels de déformation prosodique sont généralement suffisamment puissants pour pouvoir tout produire : des formes sensées et discernables, comme des non sensées et indiscernables. Mais qu'importe une formalité sans rapport avec le réel de la communication humaine ? L'artisan de la prosodie, qui œuvre sur du sens expressif, doit par conséquent rester vigilant pour garantir le sens en sortie : il doit veiller à ce que les formes issues des transformations formelles s'inscrivent dans la société des formes expressives discernables, acceptables et partageables. C'est la raison de l'approche que nous suivons en synthèse vocale : la tolérance dans l'acceptabilité expressive doit constamment trouver son fondement et sa mesure dans un corpus de formes expressives initial, constitué dans des conditions socialement normées.

4. L'expressivité dans la synthèse vocale

Notre analyse était motivée par des impératifs applicatifs, nous ne l'avions pas caché. Elle prend appui sur une réflexion portant sur l'expressivité orale en aspirant à rendre la voix de synthèse mieux recevable, puisque précisément respectueuse des conditions et des normes de l'expression humaine. Elle est ainsi aux origines d'un paradigme méthodologique basé sur une approche de la synthèse par corpus, que nous résumons pour clore (Suciu et al. 2006, Kanellos et al. 2007).

Nos précédentes discussions montrent au moins la nécessité de données représentant la norme expressive. Nous la cherchons tout d'abord dans un corpus de textes, lus dans des conditions contrôlées.

En choisissant un genre de texte particulier (p.e. conte, horoscope, nouvelles…), nous faisons varier la situation de locution pour obtenir des variantes de lecture qui correspondent à des types discursifs divers, attestés dans des pratiques socialement reconnues (discours politique, retransmission d'un match de foot, sermon…). Une fois calibré par un traitement de signal adéquat, ce corpus fait ensuite l'objet d'une entreprise de formalisation, consistant à décrire les formes expressives identifiées aux niveaux d'analyse que nous avions choisis (*syl*, *syn* et *gph*).

Dans un deuxième temps, il s'agit de développer un modèle de déformation, *i.e.* un socle d'acceptabilité des tolérances expressives. L'idée est d'augmenter cette première base de données expressives, en se donnant des moyens de déformation contrôlée qui génèrent de nouvelles formes expressives, également acceptables. Les opérateurs de déformation sont ici classiques (compressions, étirements, mises en échelle etc.) ; l'importance réside moins sur leur nature et leur fonction que sur la conservation de la norme expressive. Et, discrètement, sur la capacité des nouvelles formes de conserver les isotopies expressives, de préserver la cohérence discursive et de garantir les rythmes expressifs sur un plan macroscopique.

Enfin, c'est l'heure de « l'atelier de création » de nouvelles formes expressives. Il ne s'agit toutefois pas d'improviser, mais d'appliquer les formes déjà disponibles (d'origine ou après déformation) à des textes nouveaux, relevant, bien entendu, du même genre (de nouveaux contes, de nouveaux horoscopes etc.) et de la même situation discursive. Cette démarche vise aussi à démontrer, secondairement, le caractère indissociable entre *fde* et discours (genre textuel et situation discursive) et à mettre aussi en évidence les interdépendances entre isotopies expressives et cohérence discursive. Et, partant, la sensibilité des stratégies d'interprétation qui en dérivent aux variations expressives.

5. Conclusion

Le discours oral n'est pas un être composé de deux êtres antérieurs et autonomes (le texte et l'expressivité orale) mais un être à la fois premier et complet que l'on ne saurait décomposer qu'*a posteriori* et logiquement. La synthèse vocale, à l'assaut de services concrets mais contraints dans le calculable, suit obligatoirement ce même chemin logique et inverse pour penser ses réalisations. Certes, tout texte synthétisé oralement est expressif, même le plus fade, puisqu'il ne peut échapper au cercle expressif. Mais la synthèse vocale est grandement concernée par le retour à cet être premier et complet, fait d'enchaînements et de cohérences, ce discours qui, s'il est expressif,

l'est nécessairement dans les conditions d'une réception socialement normée. Ainsi, l'expressivité y devient-elle le pendule conceptuel indispensable qui la porte tantôt vers l'étude des normes sémiotiques inhérentes aux genres de discours, aux situations discursives et aux spécificités idiolectales, tantôt vers l'analyse de la prosodie, paradis formel de toutes les ressources expressives disponibles dans une langue. Il lui importe de comprendre le rapport au texte de ces ressources espérant y puiser ce supplément d'âme, somme toute d'expressivité, qui lui est indispensable pour réaliser des synthèses de voix obéissant aux mêmes normes que les authentiques discours humains.

Bibliographie

ABITBOL J. (2005), *L'odyssée de la voix*, Laffont.

CAELEN-HAUMONT G., « Prosodie et sens : une approche expérimentale », www.revue-texto.net/Parutions/Livres-E/ Caelen/Caelen_Prosodie.html

DÉSESQUELLES A.-C. (1999), *L'expression musicale*, Publications Universitaires Européennes.

HIRST D.J. et al. (2000), « Levels of representation and levels of analysis for the description of intonation systems », in M. Horne (éd.), *Prosody : Theory and experiment*, Kluwer Academic Publishers.

KANELLOS I. et al. (2007), « Émotions et genres de locution. La reconstitution automatique du pathos en synthèse vocale », in M. Rinn (éd.), *Le Pathos en action*, Presses Universitaires de Rennes.

LACHERET-DUJOUR A. (2000), « Prosodie : niveaux d'analyse et problèmes de représentation », in Ph. Escudier et al. (eds.), *La parole*, Hermès, 245-281.

LAFONT R. (1994), *Il y a quelqu'un. La parole et le corps*, éditions Praxiling, Université P. Valéry, Montpellier III.

RASTIER F. (1987), *Sémantique interprétative*, PUF.

RASTIER F. (2001), *Arts et sciences du texte*, PUF.

SAUVANET P. (2000), *Le rythme et la raison*, vol. 1, *Rythmologiques*, Kimé, Paris.

SUCIU I. et al. (2006), « Formal expressive indiscernibility underlying a prosodic deformation model », Actes du colloque *ExLing*, 229-232, Athens.