Speech Processing et SEO : l'avenir du référencement audio (1ère partie)

Par Yann Sauvageon

Domaine :	Recherche	**Référencement**
Niveau :	**Pour tous**	Avancé

Le référencement audio est l'un des grands sujets du SEO dans les années qui viennent, avec l'omniprésence de plus en plus forte de la vidéo dans notre vie numérique. au programme de cet article en deux parties : chiffres clés sur la vidéo, historique du Speech Processing chez Google, les dernières avancées en date et notamment le Deep Learning, ainsi, bien sûr, que les applications opérationnelles en SEO aujourd'hui....

Video is BIG !

L'usage vidéo en hyper croissance

Démocratisation du haut débit, caméra HD embarquée sur les smartphones, diffusion instantanée et quasi gratuite auprès du monde entier, il n'a jamais été aussi simple de produire de la vidéo et nous n'avons jamais eu accès à autant de contenus à ce format dans l'histoire de l'humanité !

La vidéo est devenue un des contenus majeurs du web : Streaming, Catch up, Replay, MOOC, Hangout, Meerkat, Periscope… La vidéo envahit la sphère Internet et bouleverse les *business models*.

Les sites d'actualités (issus de la presse papier) investissent pour leur part massivement dans la vidéo à l'image du Wall Street Journal qui a constitué une équipe de 40 personnes exclusivement dédiées à la production de vidéos (http://digiday.com/publishers/wsj-now-40-person-video-team-cranking-40-videos-day/) et qui référence plus de 80 000 vidéos sur son site.

YouTube, quant à lui, devient un véritable concurrent pour les investissements télé. Première mondiale, l'Allemagne va prochainement intégrer YouTube dans ses calculs d'Audimat (http://www.strategies.fr/actualites/medias/1013584W/l-allemagne-va-integrer-youtube-dans-son-audimat.html). Autant dire que ça doit couiner du côté des chaines de télévision historiques qui doivent voir cela d'un mauvais œil. La vague vidéo s'annonce donc énorme !

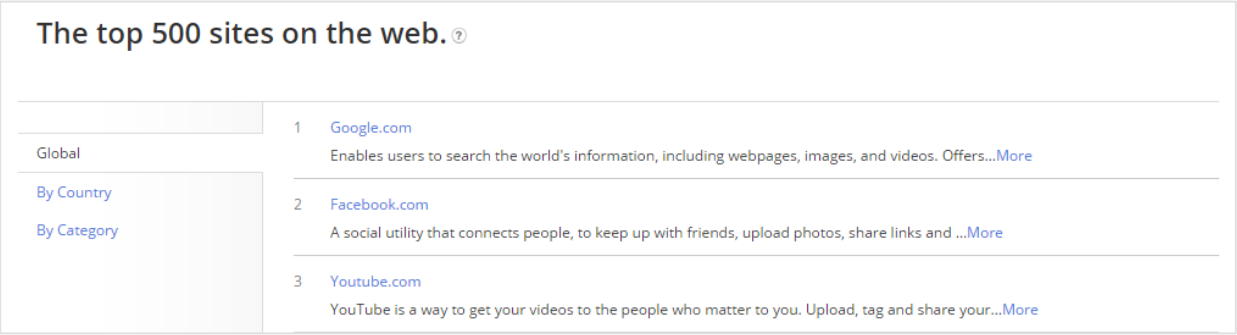

Fig.1. Classement des sites web selon Alexa : http://www.alexa.com/topsites.

COMMUNIQUÉ DE PRESSE mediametrie // NetRatings

L'audience de la vidéo sur internet en France en janvier 2015

Le Top 15 des Brands Players les plus visitées en France

Rang	Brands Player BP	Visiteurs Uniques par mois	Videos vues par mois	Total temps passé par mois en heures
1	Google/YouTube	27 014 000	1 351 606 000	41 862 000
2	Dailymotion - TS	19 670 000	298 874 000	20 883 000
3	Facebook	14 881 000	367 675 000	29 128 000
4	Digiteka - TS	11 955 000	42 357 000	1 060 000
5	TF1 / Wat - TS	11 684 000	54 241 000	10 616 000
6	France Televisions - T	7 876 000	30 016 000	7 061 000
7	CANAL+ - TS	6 652 000	24 250 000	6 239 000
8	auFeminin - TS	6 590 000	8 671 000	190 000
9	Tele Loisirs	5 654 000	66 913 000	1 698 000
10	AlloCine - TS	5 095 000	18 393 000	547 000

Fig.2. Classement des sites web selon Mediametrie/NetRatings : http://www.mediametrie.fr/internet/communiques/l-audience-de-la-video-sur-internet-en-france-en-janvier-2015.php.

Les chiffres sont vertigineux

Commençons par YouTube, fer de lance de la révolution vidéo, le site qui a récemment soufflé ses 10 bougies, concentre, à lui seul, une grande partie des vidéos de la planète web. Et pour cause, YouTube est le 2ème moteur de recherche au monde (figure 1).

Il est également, sans grande surprise, le premier lieu de consommation vidéo online en France selon Mediametrie//NetRatings (figure 2).

Les « success stories » YouTube ne manquent pas et les chiffres font tourner la tête, soit parce qu'ils dépassent l'entendement (comme « Gangnam Style » avec ses plus de 2,3 milliards de vues) soit parce qu'ils nous laissent perplexes (comme le « Nyan Cat 10 hours » avec ses plus de 40 millions de vues ?!! – posté par un certain TehN1ppe qui s'est spécialisé dans les vidéos au format « 10 hours » …).

Mais YouTube n'est que la face visible de l'iceberg Vidéo. En effet, en grande partie, les vidéos ne sont pas hébergées sur YouTube et sont diffusées par des players propriétaires, ou par d'autres solutions vidéo (Dailymotion, Brightcove, Kaltura, UStream…). Ces vidéos représentent une richesse de contenus inégalée. Les MOOC (*Massive Online Open Courses*) comme Coursera ou FUN (*France Université Numérique*) en France sont par exemple des contenus éducatifs accessibles à tous, et dont, pour l'instant, les contenus vidéos ne sont pas crawlés par Google.

Ça va s'accélérer

Les chiffres sont certes vertigineux mais l'histoire ne fait que commencer. L'usage vidéo devrait croître à un rythme très soutenu sur les 10 années à venir. D'ici 3 ans, en 2018, Cisco prédit que la vidéo représentera 80% du trafic internet mondial.

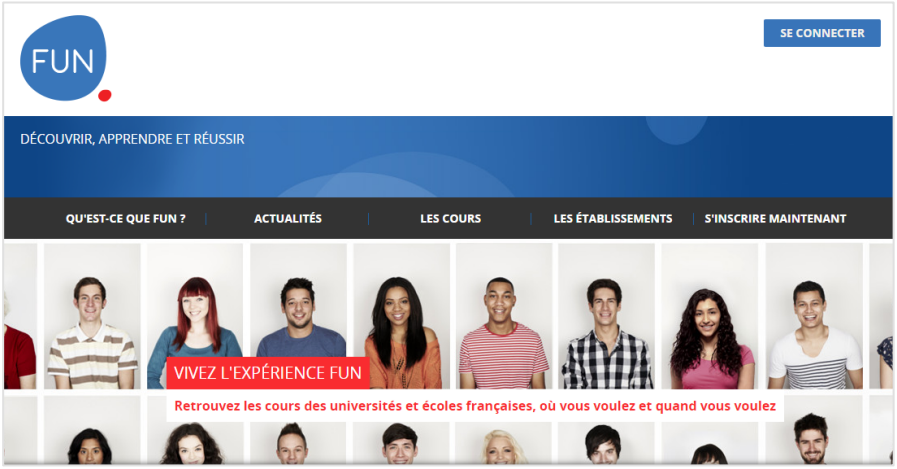

Fig.3. Le site de France Université Numérique (http://www.france-universite-numerique.fr/).

Plusieurs phénomènes vont permettre cette accélération :

• Dans un premier temps, la démocratisation de la consommation de vidéo en mobilité, notamment avec le développement conjoint du réseau 4G et de l'équipement des ménages en smartphones 4G. D'ici 2019, Cisco prédit également que près des trois quarts des données échangées sur mobile seront de la vidéo (http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html) ;

• Dans un second temps, l'arrivée de la 5G à partir de 2020. Avec la 5G, les débits théoriques en mobilité vont littéralement exploser, multipliés par 100 par rapport à ce que nous connaissons en 4G. On pourra télécharger un film de 2H en moins de 5 secondes sur son smartphone. Mais qu'à cela ne tienne, pourquoi télécharger… puisqu'on pourra streamer directement de la 4K sur son mobile !

• Et dans un troisième temps, l'omniprésence de la vidéo : Hangout, Periscope, Meerkat… la vidéo sera partout en live, en instantanée, sans coût et sans préparation, pour le meilleur et pour le pire. Les cameras seront embarquées dans les wearables et se feront très (voire trop) discrètes.

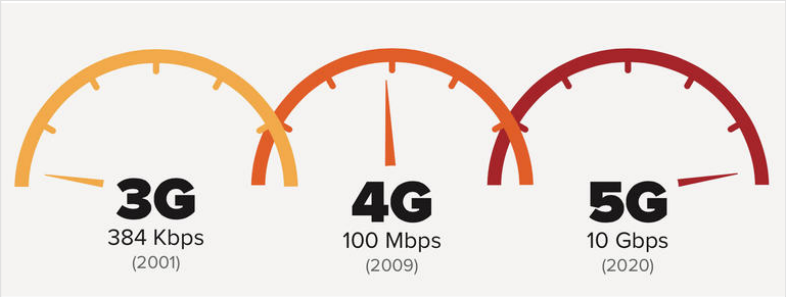

Fig.4. De la 3G à la 5G pour bientôt...

Google ne peut pas passer à côté de la vidéo

Vous l'avez compris, l'usage vidéo est majeur et le sera encore plus demain. Et il est clair que Google ne peut pas passer à côté de ces contenus. La mission de Google n'est-elle pas de « rendre toute l'information accessible de manière universelle » (http://www.google.com/about/) ?

La vidéo (et en particulier l'audio qui s'y loge) représente un contenu aujourd'hui encore peu exploité par Google. Google n'indexe pas, à proprement parler, les contenus vidéos, mais s'appuie, à l'heure actuelle, essentiellement sur les données concomitantes à la vidéo (balises, contenus textes, titres + les données d'usage sur YouTube) pour indexer et positionner les vidéos. Mais, il ne scanne pas

(encore ou en tout cas pas officiellement) les contenus audio des vidéos pour les indexer et les ranker.

Et pourtant, depuis au moins 2008, Google s'intéresse de près à ce chantier avec des améliorations majeures.

Speech Processing et Google

Juillet 2008 > Google Elections Video Search gadget

Une des premières initiatives dans le domaine du « speech to text » sur laquelle Google a largement communiqué date de 2008 et se place dans le contexte des élections présidentielles américaines.

En juillet de cette année, la Speech Team (une cellule du département Research de Google) publie le « *Google Elections Video Search gadget* ». Ce gadget permet de visionner les vidéos des candidats avec un transcript automatique. Il est également possible de rechercher directement au sein du transcript et ainsi par un système d'ancres d'accéder aux parties de la vidéo qui nous intéressent.

En limitant le champ d'application de sa technologie « speech to Text » à un nombre restreint de personnes identifiées (en l'occurrence Obama, McCain et une douzaine de candidats additionnels), Google minimise les erreurs de transcript et offre un service relativement efficace. Mais, comme Google le souligne dans son article de 2008 (http://googleblog.blogspot.fr/2008/07/in-their-own-words-political-videos.html), ce travail de « speech recognition » reste un problème difficile qui n'a pas été encore complétement résolu. (« *Speech recognition is a difficult problem that hasn't yet been completely solved, but we're constantly working to refine our algorithms and improve the accuracy and relevance of these transcribed results.* »).

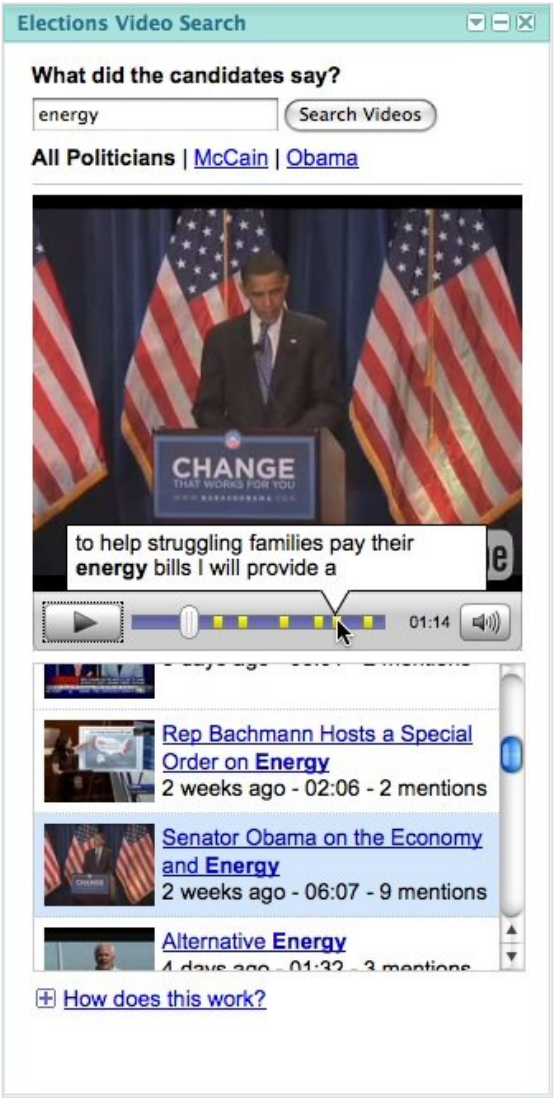

Fig.5. Le « video search gadget » de Google (2008).

Mars 2009 > Google Voice

En mars 2009, suite au rachat de GrandCentral, Google, encore une fois, sort un nouveau service de téléphonie nommé Google Voice. Ce service intègre, alors, une fonctionnalité de « Voice Transcription » dont la promesse est de fournir à ses utilisateurs un transcript automatique de ses messages vocaux en messages texte. Le champ

Fig.6. Google Voice (2009).

d'application de la technologie « speech to text » s'élargit ici avec des contextes très variés (locuteurs, signature vocale, bruit de fond…).

Le service, bien qu'innovant, est alors loin d'être 100% fiable, à l'image de ce père de famille qui, alors qu'il est chez DisneyLand, voit le message vocal de sa femme «j'ai eu les pass rapides » retranscrit en « en fait, bastard » (figure 6) ! L'histoire ne dit pas comment il a réglé l'histoire avec sa femme…

Novembre 2009 à Novembre 2012 > Transcript Automatique YouTube

Les choses sérieuses débutent en novembre 2009 avec l'activation des transcripts automatiques sur YouTube, en anglais dans un premier temps, puis dans plusieurs autres langues, dont le Français, en novembre 2012. Ce sont, au total, 10 langues qui bénéficient actuellement du transcript automatique YouTube : Allemand, Anglais, Coréen, Espagnol, Français, Italien, Japonais, Néerlandais, Portugais, Russe. Mais, là encore, malgré un déploiement à grande échelle, un simple test, au moins sur le Français, montre que le transcript est souvent bien loin de la réalité de la parole. Voir figure 7 : le résultat est loin d'être parfait.

Fig.7. YouTube a parfois du mal avec les traductions automatiques. OK, c'est Jean-Claude Van Damme, mais quand même...

Comme le souligne Brad Ellis, YouTube Product Manager, lors d'une conférence donnée fin 2013 (https://www.youtube.com/watch?v=1DFKl3MNUHs), Google sait qu'il reste des problèmes à régler. Mais compte-tenu de leur objectif à très long terme de rendre chaque vidéo compréhensible par chaque utilisateur, la technologie est le seul moyen que Google a à sa disposition. En d'autres termes, Google ne peut pas compter sur un transcript manuel à grande échelle et c'est par des technologies de « speech to text » de plus en plus performantes qu'ils y arriveront. D'ailleurs, conscient de l'ampleur de la mission, durant cette conférence, Brad Ellis laisse penser qu'il leur faudra probablement 20 à 50 ans pour espérer arriver à un résultat proche du 100% fiable.

Un département de recherche dédié au « Speech Processing »

Pour arriver à ses fins, Google a mis en place au sein de son département Research une cellule dédiée au « Speech Processing » dont l'objectif affiché est clair : rendre chaque vidéo sur le web accessible et « recherchable ». Les recherches de Google sur ce sujet sont soutenues comme le témoigne le rythme des publications sur la page suivante : http://research.google.com/pubs/SpeechProcessing.html.

Plus de concurrence aujourd'hui

Google n'est pas seul dans la course

Bien sûr, Google n'est pas seul à se battre sur le front du « speech processing » appliqué au Search.
Microsoft a notamment dans son arsenal le programme MAVIS pour « *The Microsoft Audio Video Indexing Service* » qui est un produit abouti et commercialisé. Des sociétés françaises comme Vocapia se démarquent également en étant à la pointe des technologies de Speech Processing. Sans compter IBM, Baidu, Apple, ou encore Facebook qui vient d'acquérir début 2015 la startup Wit.ai spécialisée en « Speech Recognition »... Pour n'en citer que quelques-uns !

Mais encore des limitations...

Malgré tout ce beau monde, il existe encore des limitations à la pertinence du « Speech to Text ». Ainsi, lorsqu'on nous annonce des taux de précision de 95%, on oublie de nous préciser qu'il s'agit d'un taux au niveau d'un mot clé. Ainsi, un taux de précision de 95% (au niveau du mot clé) correspond à un taux de précision de seulement 60% sur un texte de 10 mots !

Le speech processing a ses limites car il cumule plusieurs étapes probabilistes : l'identification des phonèmes dans un premier temps, puis la reconstitution de mots à partir de ses phonèmes et au final la livraison d'une phrase ayant du sens. A chaque étape doit être fait un choix entre plusieurs alternatives multipliant ainsi les sources d'erreurs.

Taux de précision mot clé	Mots clés incorrects > 1 sur x	Précision sur un texte de 8 mots	Précision sur un texte de 10 mots
50%	1 sur 2	0%	0%
67%	1 sur 3	4%	2%
75%	1 sur 4	10%	6%
85%	1 sur 7	27%	20%
90%	1 sur 10	43%	35%
95%	1 sur 20	66%	60%
99%	1 sur 100	92%	90%

Fig.8. Taux de précision de la reconnaissance vocale en fonction de la taille du texte.

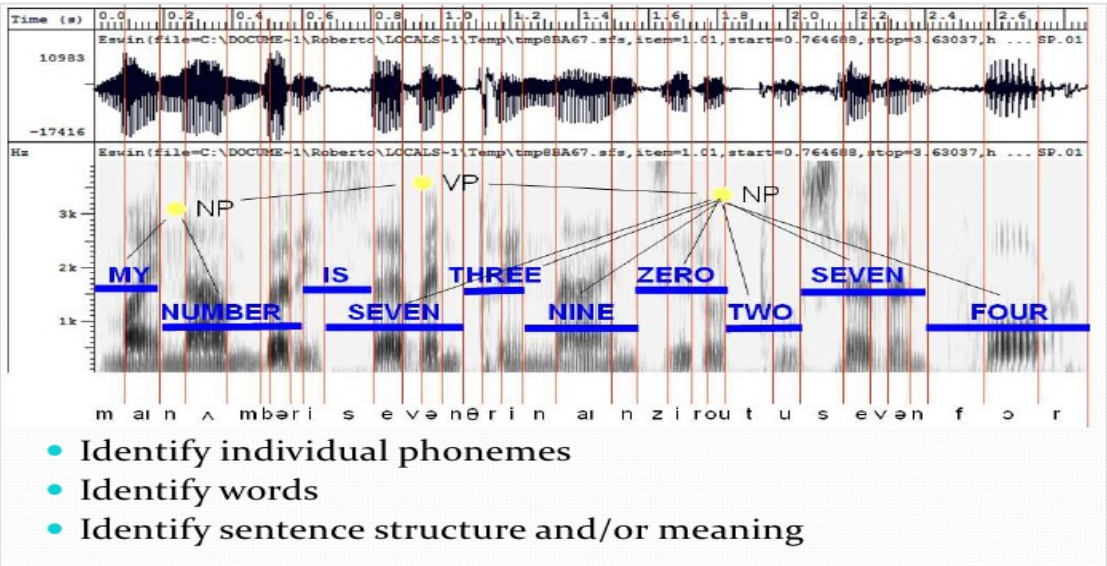

Fig.9. Identification des phonèmes d'une phrase.

Le mois prochain, nous continuerons cet article avec la notion, aujourd'hui essentielle, le *deep learning*, et ses promesses dans le domaine de la reconnaissance vocale. Nous envisagerons également le référencement audio et les différentes stratégies à mettre en place à ce niveau pour les années qui viennent. A très bientôt !

Yann Sauvageon,
Directeur de l'expertise, Synodiance (http://www.synodiance.com/).

Cet article fait suite à la conférence co-animée par l'auteur avec Jérôme Rocheteau au SEO Campus 2015, dont vous trouverez les slides ici :
http://www.slideshare.net/Synodiance/synodiance-seo-et-speech-processing-futur-enjeu-seo-seo-campus-2015